

# A genomics approach to the conservation of kōwaro (Canterbury mudfish, *Neochanna burrowsius*)

A thesis submitted in partial fulfilment

of the requirements for the degree of

Master of Science in Biology

at the

University of Canterbury

by

Levi Fayne Maakawhio Collier-Robinson

University of Canterbury

School of Biological Sciences

2019

## Acknowledgements

Mokopiki, mokokake

Piki ake Tāwhaki ki te raki tuatahi

Kake ake Tāwhaki ki te raki tuarua

Haere ake Tāwhaki ki te raki tuakāhuru

Ka puta kai ruka ki te hārorerore

Ka puta ki takata okotahi

Ki a Rehua e!

Tīhei, mauri ora!

Ko te kāhui mauka, tū tonu, tū tonu. Ko te kāhui takata, karo noa, karo noa, ka haere. Koutou kā manu pīrau a Tāne, moe mai rā, okioki mai rā, kia kore e warewaretia.

Āpiti hono, tātai hono, rātou ki a rātou, ka huri nei ki a tātou, te huka ora.

Ka rere tāku manu ki te Tai o Poutini ki te mihi ki kā wai kōratarata o Makawhio, o Arahura, o Kawatiri hoki. Ka whakawhiti i Kā Tiritiri o te Moana ki Kā Pākihi Whakatekateka o Waitaha, kia noho ai ki tōku papa kāika o Tuahiwi.

Tiro ake ana ki te kāhui whetū o Rākaihautū e tū mai rā, nāhana i tīmata te ahi tuatahi ki tēnei whenua. Ko Apa, ko Waitaha, ko Māmoe, ko Tahu, ko au.

Ka mihi ki ōku kaiarahi i tautoko i aku nei mahi. Tuatahi ki a Tammy Steeves, tōku pouako matua mo tēnei mahi rangahau, nei rā te mihi matakaui matakoro ki a koe mō āu kōrero akiaki kia whaia tēnei iti kahuraki. Ka nui rawa kā mihi ki a koe mo te tākohataka o ōu mōhiotaka katoa. Tuarua ki a Angus McIntosh, ko koe te take kei te ako tonu au. Kā mihi maioha rawa ki a koe mō tō whakatenatena i au ki te noho ki te whare wānaka, kia whai tonu i tēnei tohu. Ki ōku hoa mahi katoa o kā rōpu whakaharahara o te kura pūtaiao koiora, arā, ki kā rōpu o CONSert rāua ko FERG, tēnā koutou katoa. Ki te wero pūtaiao o ngā koiora tuku iho o Aotearoa, tēnā rā koutou katoa mō te

pūtea me kā whakawhitinga kōrero. Ka mihi hoki ki NTRC mō te tautoko i āku mahi katoa ki Te Whare Wānaga o Waitaha. E kore rawa āku mihi e mutu mō ōku mātua. Kua tautoko kōrua i ōku akoka katoa, ahakoa kā piki me kā heke. Kia mutu, ka mihi hoki ki te kaipānui, tēnā koe.

Ānei rā he mahi kaitakata. Inaianei, kua kakea kētia Kā Tiritiri o te Moana, ki tua he pākihi rauarahi, he whenua haumako muia e te takata.

## Abstract

Often the most appropriate sampling strategies and genomic approaches to informing the conservation management of threatened taonga species in Aotearoa New Zealand in a culturally responsive manner is unclear. Here I investigate this, using the critically endangered kōwaro (Canterbury mudfish; *Neochanna burrowsius*) as a case study. Firstly, by examining the effect of sample size on measures of genetic diversity to determine a cost-effective approach to sampling threatened taonga species like kōwaro for population genomics research.

In Te Ao Māori, genomic data obtained from taonga species is tapu and best studied using kaupapa Māori principles. To achieve this, my co-authors and I co-developed a research programme with Ngāi Tūāhuriri that integrates kaupapa Māori with emerging genomic technologies and extensive ecological data for two taonga species. Chapter Three outlines this broader research programme, the foundation of which, is an iterative decision-making framework of critical steps in genomic research that includes tissue sampling as well as data generation, storage and access and how responsiveness at each of these stages encourages the expression of Māoritanga.

## Table of Contents

Acknowledgements .....	2
Abstract .....	4
Table of Contents .....	5
Chapter One: Introduction.....	6
Conservation genomics.....	6
Kōwaro.....	8
Sampling strategies .....	10
Thesis structure and chapter outlines .....	12
References .....	13
Chapter Two: How many samples are necessary for population genomic research on threatened species? .....	18
Introduction .....	18
Methods.....	20
Results.....	23
Discussion .....	29
References .....	32
Chapter Three: Embedding kaupapa Māori principles in genomic research of taonga species: a conservation genomics case study.....	36
Abstract.....	36
Ngā taonga tuku iho .....	40
Key kaupapa Māori principles for genomic research on taonga species.....	43
Genomic research.....	45
Case study .....	47
References .....	53
Chapter Four: Discussion .....	60
Conservation genomics in Aotearoa New Zealand.....	60
Future directions for the genomics research of kōwaro .....	61
References .....	62
Appendix A: Scripts for generating resampled datasets .....	65
Appendix B: R code for resampled datasets .....	71

# Chapter One: Introduction

## Conservation genomics

Early Māori and European settlers have both had a significant negative impact on the ecology of many New Zealand ecosystems (Saunders and Norton 2001). The conservation of our native and endemic species is necessary to combat the loss of biodiversity that New Zealand is experiencing (Craig et al. 2000). Many traditional Māori cultural values and ideas align well with modern conservation (Roberts et al. 1995; Kawharu 2000), and there is increasing involvement of local iwi and hapū in the conservation and co-management of threatened taonga species. Some key strategies that are being implemented to conserve biodiversity include ecological restoration (e.g. Saunders and Norton 2001) and species translocations (e.g. Griffith et al. 1989; Miskelly and Powlesland 2013). Translocation is defined as the deliberate release of organisms to establish, re-establish or augment populations (Griffith et al. 1989; IUCN/SSC 2013). It has been a very successful management strategy for the conservation of a range of native taxa in New Zealand and elsewhere (Griffith et al. 1989; Towns and Ferreira 2001; Miskelly and Powlesland 2013).

Conservation management practices, including translocation, are continuously improving to enable more successful conservation of threatened species and ecosystems. There are a range of risks involved with translocations which has led to the development of guidelines that are designed to maximize success (IUCN/SSC 2013). Among these risks are a range of genetic issues that must be addressed to enable the long-term success of translocations. Many species of conservation concern have small, fragmented populations that can suffer from detrimental genetic effects. These include the stochastic loss of genetic diversity due to genetic drift (Frankham 2010) and the expression of deleterious recessive alleles due to inbreeding (inbreeding depression), which reduces fitness (Keller and Waller 2002; Grueber et al. 2010). When translocating individuals to establish a new population, it is important to be mindful of these issues that can arise from low genetic diversity. In addition to the short-term effects (i.e. reduced fitness), low genetic diversity can have long-term effects by

limiting the populations ability to adapt to changes in their environment (adaptive potential, REF).

Translocations to augment small populations can also alleviate low genetic diversity by increasing the effective population size and introducing new alleles (Whiteley et al. 2015), however, there can be a risk of outbreeding depression (Frankham et al. 2011; but see Ralls et al. 2017).

Genetic and, more recently, genomic approaches are being increasingly used to inform conservation strategies, including translocations (Weeks et al. 2011; Funk et al. 2012). In contrast to traditional genetic techniques that used a small set of genetic markers scattered across the genome (Frankham et al. 2010), high-throughput DNA sequencing enables the generation of many thousands of single nucleotide polymorphisms (SNPs) to inform conservation management strategies (Allendorf et al. 2010; Funk et al. 2012). *De novo* approaches can be used to genotype individuals but developing appropriate genomic resources such as a reference genome can greatly increase the likelihood of detecting SNPs by enabling a reference guided approach to genotyping (Davey et al. 2011; Elshire et al. 2011; Torkamaneh et al. 2016). Genotyping-by-Sequencing (GBS) is an approach that uses restriction enzymes to cut genomic DNA in to fragments, and these fragments are then used to prepare a GBS library for high-throughput sequencing (Elshire et al. 2011). GBS is a cost-effective method that takes advantage of reduced representation for the simultaneous discovery and genotyping of thousands of genome-wide SNPs (Elshire et al. 2011; Torkamaneh et al. 2016). Although more expensive than GBS, whole genome resequencing can greatly expand on the potential of reduced representation approaches by detecting more SNPs at more loci throughout the genome (i.e., by detecting SNPs not associated with restriction sites).

Deciding which strategy will be the most effective for species with no genomic resources available (and very few from closely related species) can be challenging. This is the reality for many threatened endemic species, including kōwaro (Canterbury mudfish; *Neochanna burrowsius*), a critically endangered freshwater fish species (Goodman et al. 2014; Dunn et al. 2018).

## Kōwaro

Galaxiids in New Zealand face a range of threats such as predation from exotic fishes (McIntosh et al. 2010), and kōwaro are no exception. There have been many observations of considerable predation of kōwaro by trout and eels (e.g. Eldon 1979). Co-occurrence of kōwaro and their predators is therefore very low, due to competitive exclusion (O'Brien and Dunn 2007). However, *Neochanna* sp. exhibit a range of physiological and behavioural adaptations suited to surviving extreme conditions that would be lethal to many other fish species (Urbina et al. 2014, O'Brien and Dunn 2007). There is a significant trade-off between their biotic and abiotic tolerance, but the ability to tolerate harsh abiotic conditions in drying streams and ponds expands the realised niche of kōwaro and enables the persistence of some key populations in the absence of predators and competitors (Harding et al. 2007). These characteristics of kōwaro in combination with intensive land use change across Canterbury have led to a fragmented distribution of small populations across their range (Figure 1.1), increasing the likelihood of inbreeding depression and genetic drift exposing deleterious alleles.

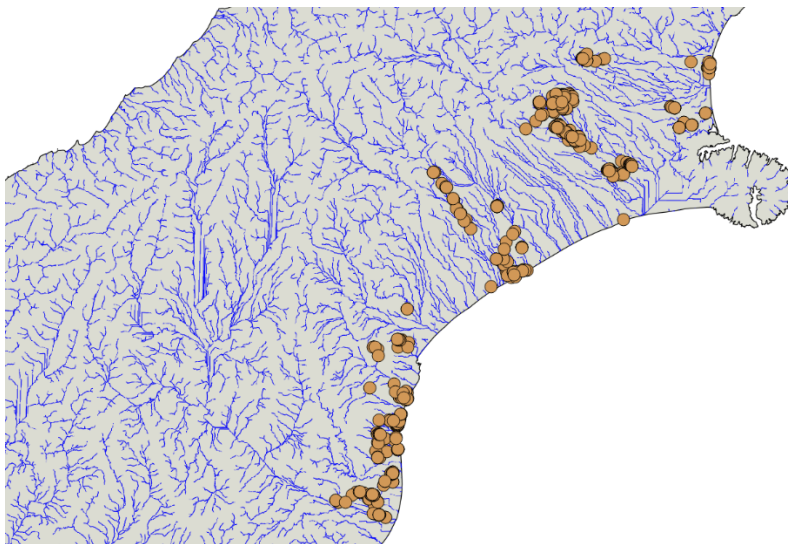


Figure 1.1: Map of all recorded kōwaro observations from the New Zealand Freshwater Fish Database (NZFFD 2019), indicating the distribution of the species within Canterbury.



Drying of freshwater habitats results in a decrease in habitat size, diversity and availability (Rolls et al., 2012). As these habitats constrict, the density of mobile aquatic animals can increase as they become concentrated in pools of decreasing size, causing increased competition due to higher encounter rates (Lake, 2003). Environmental change has the potential to increase the duration, frequency, intensity and timing of drying in freshwater habitats (Rolls et al. 2012). Although kōwaro are well equipped with adaptations suited to survive temporary drying (Meredith 1982; O'Brien and Dunn 2007, Urbina et al. 2014), extended periods of drought are known to result in high mortality and large population declines because of both physiological and ecological stressors (Eldon 1979; Urbina et al. 2014; Meijer et al. 2019). These increasingly extreme abiotic and biotic conditions can lead to local extirpation. The fragmented distribution of kōwaro then prevents the natural recolonization of isolated populations contributing to the rapid decline of the species (Dunn et al. 2018).

Increasing the connectivity between isolated kōwaro populations could allow for recolonization after local extirpations and enable gene flow to between previously isolated populations. However, this solution risks facilitating the introduction and spread of predators throughout remaining kōwaro populations, and thus lead to further declines. Therefore, translocations to new suitable locations within their range have become an important conservation management strategy to enhance species recovery, but many translocations have been unsuccessful (Barrier 2003; O'Brien and Dunn 2007). The reasons for many failed translocation attempts are largely unknown, but there has been little consideration to the number of individuals translocated as well as limited monitoring post-translocation. Therefore, there is much to improve in this space and the inclusion of genomics is a strong starting point. A conservation genomics approach to informing future translocations has the potential to improve their success rate by using measures of population genetic structure and

genomic diversity to mitigate the loss of diversity and an increase in inbreeding in the short-term, while maximizing adaptive potential in the long-term (Weeks et al. 2011).

### Sampling strategies

To date, research on kōwaro has focused on assessing population distribution, abundance and size structure (e.g. Harding et al. 2007; Meijer et al. 2019). There has been limited research on the population genetics of kōwaro. Using a single mitochondrial locus, Davey et al. (2003) found two haplotype groups that corresponded with the geographic distribution of kōwaro (Figure 1.2). These data led to the recommendation that the Northern and Southern populations of kōwaro be managed as separate units (Figure. 2; Davey et al. 2003; Barrier 2003; O'Brien and Dunn 2005; O'Brien and Dunn 2007) and have since been used in part to identify “key” kōwaro populations to be conserved (Barrier 2003, O'Brien and Dunn 2007). Conservation genomics present an opportunity to not only re-examine these conservation units for kōwaro but also to expand our understanding of their genetic diversity, both within and between populations, to inform the future conservation and management of this taonga species.

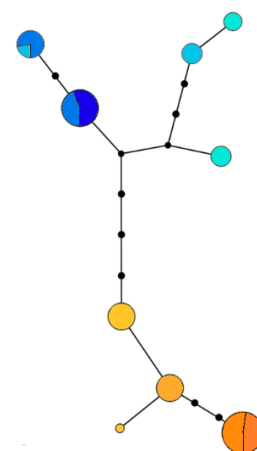
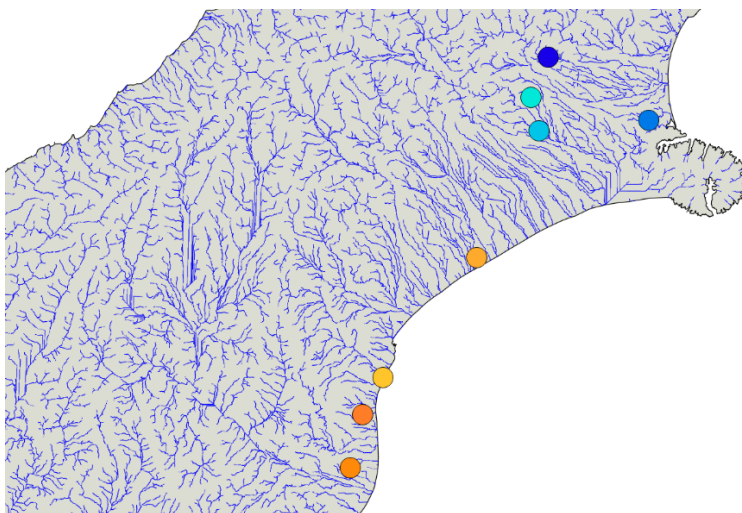


Figure 1.2: a) Map of kōwaro populations sampled in Davey et al. (2003). Blue circles denote 'Northern' populations and orange circles 'Southern' populations. b) Haplotype network for kōwaro mitochondrial control region haplotypes generated using TCS (Clement 2002) and redrawn using PopArt (Leigh and Bryant 2015), showing population genetic structure between 'Northern' and 'Southern' populations. Circle sizes proportional to haplotype frequency. Black circles represent missing haplotypes. (redrawn from Davey et al. 2003).

Davey et al. (2003) provides an excellent starting point for locating appropriate sampling locations to assess kōwaro diversity. However, there is still no genetic information available for most of the populations across their range (Figure 1.1, Figure 1.2). Previous ecological studies have identified a range of different habitat types in which kōwaro are found, as well as threats to kōwaro such as predators, competition or drying intensity that could be divergently driving selection in different populations (e.g. Eldon 1979; Harding et al. 2007; Meijer et al. 2019). Sampling a range of populations in diverse habitats can enable more robust estimates of the overall genetic diversity found between populations (Hoban et al. 2016).

Another challenge in sampling kōwaro or any other species is the question of how many samples are needed per population? When using a small number of genetic markers, relatively large sample sizes are typically recommended ( $n \geq 20$ ) (Yan and Zhang 2004; Ryman et al. 2006; Miyamoto et al. 2008; Pruett and Winter 2008; Hale et al. 2012). However, it is unclear whether these large sample sizes are necessary when using large numbers of genome-wide SNPs (Willing et al. 2012; Nazareno et al. 2017; Gaughran et al. 2018; Flesch et al. 2018). However, these studies used species with large population sizes, so there is a need to address this question for the specific purpose of the conservation management of threatened taonga species with small, isolated populations such as kōwaro.

Kōwaro are taonga to Ngāi Tahu, the predominant iwi in Te Waipounamu (South Island) and mana whenua in Canterbury. All research on taonga species can benefit from genuine co-development with local Māori. This influences the sampling strategy because including areas of cultural significance can enhance the recovery of kōwaro populations in these areas, consequently enhancing the expression of Māori values associated with the species and these places. Therefore, where appropriate, these locations should also be included in the population genomic sampling of kōwaro.

### Thesis structure and chapter outlines

I have structured my thesis around two data chapters intended to be published as stand-alone multi-authored papers led by me. This includes one manuscript that is currently in review for the *New Zealand Journal of Ecology* 2019 Special Issue entitled “Mātauranga Māori and shaping ecological features” (Chapter Three). Given the structure of my thesis, there is overlapping material between chapters and each chapter has its own references section. Although I have led the research presented in this thesis, this work is a product of collaborative efforts, which is especially relevant for Chapter Three.

*Ehara tāku toa i te toa takitahi, ēngari he toa takitini.*

Chapter Two examines the relationship between sample size and measures of genetic diversity to determine a cost-effective approach to sampling threatened taonga species like kōwaro for population genomics research.

Chapter Three outlines the collective experience of myself and my colleagues in embedding kaupapa Māori in the genomic research of taonga species, providing an iterative-decision making framework that highlights critical stages in genomic research of taonga species and how responsiveness at each of these stages encourages the expression of Māoritanga.

Chapter Four provides a summary of the research presented in my thesis and considerations for a conservation genomics approach for threatened taonga species like kōwaro, within the bicultural research environment of Aotearoa New Zealand.

The purpose of my thesis is to outline the most appropriate sampling strategies and genomic approaches to informing the conservation management of threatened taonga species in Aotearoa New Zealand, using kōwaro as a case study.

## References

- Allendorf, F.W.; Hohenlohe, P.A.; Luikart, G. 2010. Genomics and the future of conservation genetics. *Nature reviews genetics* 11: 697.
- Barrier, R. 2003. *New Zealand Mudfish (Neochanna Spp.) Recovery Plan 2003-13: Northland, Black, Brown, Canterbury, and Chatham Island Mudfish*. Department of Conservation.
- Clement, M.J., Snell, Q., Walker, P., Posada, D. and Crandall, K.A., 2002, April. TCS: estimating gene genealogies. In *ipdps* (Vol. 2, p. 184).
- Craig, J., Anderson, S., Clout, M., Creese, B., Mitchell, N., Ogden, J., Roberts, M. and Ussher, G., 2000. Conservation issues in New Zealand. *Annual Review of Ecology and Systematics*, 31(1), pp.61-78.
- Davey, M. L., O'Brien, L., Ling, N., & Gleeson, D. M. 2003. Population genetic structure of the Canterbury mudfish (*Neochanna burrowsius*): biogeography and conservation implications. *New Zealand Journal of Marine and Freshwater Research*, 37(1), 13-21.
- Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. and Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7), p.499.
- Dunn, N.R.; Allibone, R.M.; Closs, G.; Crow, S.; David, B.O.; Goodman, J.; Griffiths, M.H.; Jack, D.; Ling, N.; Waters, J.M. 2018. *Conservation status of New Zealand freshwater fishes, 2017*. Publishing Team, Department of Conversation.

Eldon, G. A., 1979. Habitat and interspecific relationships of the Canterbury mudfish, *Neochanna burrowsius* (Salmoniformes: Galaxiidae). *New Zealand journal of marine and freshwater research*, 13(1), 111-119.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, 6(5), p.e19379.

Flesch, E.P., Rotella, J.J., Thomson, J.M., Graves, T.A. and Garrott, R.A., 2018. Evaluating sample size to estimate genetic management metrics in the genomics era. *Molecular ecology resources*, 18(5), pp.1077-1091.

Frankham, R.; Ballou, J.D.; Briscoe, D.A. 2010. *Introduction to Conservation Genetics*. Cambridge University Press.

Frankham, R., Ballou, J.D., Eldridge, M.D., Lacy, R.C., Ralls, K., Dudash, M.R. and Fenster, C.B., 2011. Predicting the probability of outbreeding depression. *Conservation Biology*, 25(3), pp.465-475.

Funk, W.C.; McKay, J.K.; Hohenlohe, P.A.; Allendorf, F.W. 2012. Harnessing genomics for delineating conservation units. *Trends in ecology & evolution* 27: 489-496.

Goodman, J.M., Dunn, N.R., Ravenscroft, P.J., Allibone, R.M., Boubee, J.A.T., David, B.O., Griffiths, M., Ling, N., Hitchmough, R.A., Rolfe, J.R. 2014. *New Zealand Threat Classification Series 7*. Department of Conservation, Wellington. 12 p.

Grueber, C.E., Laws, R.J., Nakagawa, S. and Jamieson, I.G., 2010. Inbreeding depression accumulation across life-history stages of the endangered takahe. *Conservation Biology*, 24(6), pp.1617-1625.

Griffith, B., Scott, J.M., Carpenter, J.W. and Reed, C., 1989. Translocation as a species conservation tool: status and strategy. *Science*, 245(4917), pp.477-480.

Hale, M.L., Burg, T.M. and Steeves, T.E., 2012. Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PloS one*, 7(9), p.e45170.

Harding, J.S., 2007. Persistence of a significant population of rare Canterbury mudfish (*Neochanna burrowsius*) in a hydrologically isolated catchment. *New Zealand Journal of Marine and Freshwater Research*, 41, pp.309-316.

Hoban, S., Kelley, J.L., Lotterhos, K.E., Antolin, M.F., Bradburd, G., Lowry, D.B.; Poss, M.L., Reed, L.K., Storfer, A., Whitlock, M.C. 2016. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist* 188: 379-397.

IUCN/SSC 2013. Guidelines for Reintroductions and Other Conservation Translocations. Version 1.0. Gland, Switzerland: IUCN Species Survival Commission, viiii + 57 pp.

Kawharu, M. 2000. Kaitiakitanga: a Maori anthropological perspective of the Maori socio-environmental ethic of resource management. *Journal of the Polynesian Society* 109: 349-370.

Keller, L.F. and Waller, D.M., 2002. Inbreeding effects in wild populations. *Trends in ecology & evolution*, 17(5), pp.230-241.

Lake, P. S. 2003. Ecological effects of perturbation by drought in flowing waters. *Freshwater biology*, 48(7), 1161-1172.

Leigh, J.W. and Bryant D., 2015. PopART: Full-feature software for haplotype network construction. *Methods in Ecology and Evolution*, 6(9):1110–1116.

McIntosh, A.R., McHugh, P.A., Dunn, N.R., Goodman, J.M., Howard, S.W., Jellyman, P.G., O'Brien, L.K., Nyström, P. and Woodford, D.J., 2010. The impact of trout on galaxiid fishes in New Zealand. *New Zealand Journal of Ecology*, 34(1), p.195.

Meijer, C.G., Warburton, H.J., Harding, J.S. and McIntosh, A.R., 2019. Shifts in population size structure for a drying-tolerant fish in response to extreme drought. *Austral Ecology*.

Meredith, A. S., Davie, P. S., & Forster, M. E. 1982. Oxygen uptake by the skin of the Canterbury mudfish, *Neochanna burrowsius*. *New Zealand Journal of Zoology*, 9(3), 387-390.

Miskelly, C. M. and Powlesland, R. G. 2013. Conservation translocations of New Zealand birds, 1863–2012. *Notornis*, 60, 3-28.

Miyamoto, N., Fernández-Manjarrés, J.F., Morand-Prieur, M.E., Bertolino, P. and Frascaria-Lacoste, N., 2008. What sampling is needed for reliable estimations of genetic diversity in *Fraxinus excelsior* L.(Oleaceae)?. *Annals of forest science*, 65(4), p.1.

Nazareno, A.G., Bemmels, J.B., Dick, C.W. and Lohmann, L.G., 2017. Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Molecular Ecology Resources*, 17(6), pp.1136-1147.

NZFFD 2019. Retrieved from <https://www.niwa.co.nz/information-services/nz-freshwater-fish-database>

O'Brien, L. and Dunn, N. 2007. *Mudfish (Neochanna Galaxiidae) literature review*. Science & Technical Pub., Department of Conservation.

Pruett, C.L. and Winker, K., 2008. The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. *Journal of Avian Biology*, 39(2), pp.252-256.

Ralls, K., Ballou, J.D., Dudash, M.R., Eldridge, M.D., Fenster, C.B., Lacy, R.C., Sunnucks, P., Frankham, R. 2017. Call for a paradigm shift in the genetic management of fragmented populations. *Conservation Letters* 11: e12412.

Roberts, M., Norman, W., Minhinick, N., Wihongi, D., & Kirkwood, C. 1995. Kaitiakitanga: Maori perspectives on conservation. *Pacific Conservation Biology*, 2(1), 7-20.

Rolls, R. J., Leigh, C., & Sheldon, F. 2012. Mechanistic effects of low-flow hydrology on riverine ecosystems: ecological principles and consequences of alteration. *Freshwater Science*, 31(4), 1163-1186.

Ryman, N., Palm, S., André, C., Carvalho, G.R., Dahlgren, T.G., Jorde, P.E., Laikre, L., Larsson, L.C., Palmé, A. and Ruzzante, D.E., 2006. Power for detecting genetic divergence: differences between statistical methods and marker loci. *Molecular Ecology*, 15(8), pp.2031-2045.

Saunders, A. and Norton, D.A., 2001. Ecological restoration at mainland islands in New Zealand. *Biological Conservation*, 99(1), pp.109-119.

Torkamaneh, D., Laroche, J. and Belzile, F., 2016. Genome-wide SNP calling from genotyping by sequencing (GBS) data: a comparison of seven pipelines and two sequencing technologies. *PLoS One*, 11(8), p.e0161333.

Towns, D. R., and Ferreira, S. M. 2001. Conservation of New Zealand lizards (Lacertilia: Scincidae) by translocation of small populations. *Biological conservation*, 98(2), 211-222.



Urbina, M.A., Meredith, A.S., Glover, C.N. and Forster, M.E., 2014. The importance of cutaneous gas exchange during aerial and aquatic respiration in galaxiids. *Journal of Fish Biology*, 84(3), pp.759-773.

Weeks, A. R., Sgro, C. M., Young, A. G., Frankham, R., Mitchell, N. J., Miller, K. A., ... & Breed, M. F. 2011. Assessing the benefits and risks of translocations in changing environments: a genetic perspective. *Evolutionary Applications*, 4(6), 709-725.

Whiteley, A.R., Fitzpatrick, S.W., Funk, W.C. and Tallmon, D.A., 2015. Genetic rescue to the rescue. *Trends in ecology & evolution*, 30(1), pp.42-49.

Willing, E.M., Dreyer, C. and Van Oosterhout, C., 2012. Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PloS one*, 7(8), p.e42649.

Yan, L. and Zhang, D., 2004. Effects of sample size on various genetic diversity measures in population genetic study with microsatellite DNA markers. *Dong wu xue bao.[Acta zoologica Sinica]*, 50(2), pp.279-290.

## Chapter Two: How many samples are necessary for population genomic research on threatened species?

### Introduction

The recent expansion of genomics has significantly reduced sequencing costs and population genomic approaches like reduced-representation sequencing and whole genome resequencing (WGR) have never been more accessible, especially for threatened species. Reduced representation sequencing (e.g. GBS) can be more cost-effective than traditional approaches (Galla et al. 2016), especially when there are existing genomic resources for closely related species (Galla et al. 2019). While WGR is generally a more expensive approach than GBS, it is becoming a more effective approach for some species if the genome is relatively small (Galla et al. 2019). For example, in the last three years, whereas the cost of generating GBS data has remained unchanged, the cost of resequencing 10X bird genomes (1.3 Gb) has reduced by approximately two-thirds (T. Steeves, personal communication). The motivation for using WGR over reduced-representation approaches is two-fold: more data will be generated, and the data that is generated will be more representative of the entire genome (i.e., not only of those regions associated with restriction sites; Fuentes-Pardo and Ruzzante 2017). Budget limitations can be overcome by reducing the amount of WGR required. Whether that is by utilizing existing genomic resources (e.g. Galla et al. 2019), or by reducing the number of samples, or by reducing the depth of coverage at which those samples are sequenced. Thus, an evidence-based sampling design that minimizes sample size or depth of coverage may be an effective way to decrease costs and thereby enable a WGR approach for species with relatively small genomes. However, there is always the risk that reducing sample sizes and depth of coverage may result in inaccurate estimates of genome-wide diversity (e.g. if depth of coverage is too low, it becomes harder to call heterozygous SNPs). The ultimate risk of this is that inaccurate estimates of

population genomic structure or genomic diversity may negatively impact conservation management recommendations for already threatened species.

Several studies have demonstrated how many samples are necessary for population studies using traditional genetic markers (Yan and Zhang 2004; Ryman et al. 2006; Miyamoto et al. 2008; Pruett and Winter 2008; Hale et al. 2012). For example, Hale et al. (2012) demonstrated for four taxonomically diverse species with different effective population sizes that large sample sizes ( $n=25-30$  individuals) are necessary for accurate measures of allele frequencies when using small numbers of microsatellite markers ( $< 10$  markers). It is unclear whether this is also true when using many thousands of genomic markers (i.e. SNPs), particularly for threatened species with low effective population sizes. Although there is some evidence that when using large numbers of SNPs (1,000-23,000 SNPs) generated from reduced-representation approaches, smaller sample sizes ( $n = 2-10$ ) may be sufficient (Willing et al. 2012; Nazareno et al. 2017), other studies have continued to recommend large sample sizes ( $n = 25$ ) even when using large numbers of SNPs ( $\sim 12,000$  SNPs; Flesch et al. 2018). However, these studies used species non-threatened species with relatively large census sizes (and therefore likely relatively large effective populations). Thus, there is a need to address this question specifically for critically endangered species like kōwaro.

In addition to considering the effective population size of a focal species, it is also important to consider the objectives of the study. For example, if the objective is to characterize population genomic structure and estimate genome-wide diversity, then accurate measures of heterozygosity for many SNPs distributed widely across the genome will be the most informative.

In an aligned research project, a draft reference genome for kōwaro has been generated and we now know that kōwaro have a relatively small genome of  $\sim 700$  Mb (R. Moraga unpublished data). The relatively small size of the kōwaro genome provided us with a potentially cost-effective opportunity to use a WGR approach rather than a reduced-representation approach like genotyping-by-sequencing (GBS).

Here, I use a resampling approach on available genomic datasets in two New Zealand bird species to test the effect of sample size and depth of coverage on relevant measures of diversity, with the intention of applying these results to the future generation of resequencing data for kōwaro populations.

## Methods

### Proxy focal taxon

Kakī (*Himantopus novaezealandiae*) is a critically endangered wading bird that was once widespread across Aotearoa New Zealand but is now restricted to a single wild population within the Mackenzie Basin (Robertson et al. 2018). Genetic and genomic approaches have been used to inform pairing recommendations within the captive breeding programme and Genotyping-by Sequencing (GBS) and whole genome resequencing datasets are now available for this species (Galla et al. 2019). Because kakī are restricted to one small, isolated population (the total census size for kakī is < 150 adult birds), the available kakī genomic datasets provide an excellent proxy for a resampling approach to determine the number of samples per population for other threatened species that also occur in small, isolated populations such as kōwaro. The total census size for most kōwaro populations is unknown but the logic here is that both critically endangered species are likely to have relatively low effective population sizes.

Buller's albatross (*Thalassarche bulleri*) is classified as a naturally uncommon at-risk species (Robertson et al. 2018). I also resampled a GBS dataset for two genetically distinct northern and southern populations of this species (J. Wold, unpublished data) to briefly examine the effect that larger effective population size may have on sample size recommendations. The effective population size for each of the two populations of Buller's albatross has not been estimated, but the approximate census sizes for the northern and southern populations are approximately 18,000 and

15,000 breeding pairs, respectively, so it is reasonable to assume that, compared to kakī and kōwaro, the effective population sizes of each population is relatively large.

### Generating resampling datasets

The kakī resequencing dataset included 36 total individuals, consisting of both high coverage samples ( $n = 24$ ) and low coverage samples ( $n = 12$ ), with an average depth of coverage =  $17.44 \pm 6.79$  (Galla et al. 2019). The high coverage samples provide an opportunity to sample a subset of these reads to assess the effect of depth of coverage. A series of custom scripts (Moraga 2018) were used to subsample high coverage fastq files to create separate simulated datasets with reduced depth of coverage (average depth = 10 and 6). Reads from each dataset were aligned to the kakī reference genome using Bowtie2 (Langmead and Salzberg 2012). Galla et al. (2019) used BCFtools v1.9 (Li et al. 2009) to filter the full kakī resequencing dataset ( $n=36$ ) for biallelic SNPs with a minor allele frequency  $> 0.05$ , a Phred-score  $> 20$ , maximum of 10% missing data per site and pruned for linkage disequilibrium. VCFtools v1.9 (Danecek et al. 2011) was then used to thin for 1 SNP in every 150 bp. I applied these same filtering parameters to the reduced depth datasets. In addition to these filtering settings, each resequencing dataset was also filtered for different average mean depths. The full dataset was filtered for an average mean depth  $> 10$ , this was reduced to  $> 8$  in the 10x dataset and  $> 4$  in the 6x dataset. The VCFs for each of the filtered datasets were then resampled to assess the effect of sample size. I generated simulated datasets of 1,000 replicates at each sample size for the full dataset ( $n = 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34$ ) and the reduced depth datasets ( $n=2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24$ ). For each replicate, VCFtools v1.9 (Danecek et al. 2011) was used to generate random subsets of individuals.

A resampling approach was also applied to the kakī GBS dataset to compare resequencing and GBS data and examine any effect this may have on minimum recommended sample size. Each fastq file from the full kakī GBS dataset ( $n = 88$  unique individuals, average depth =  $13.73 \pm 6.53$ ) was

resampled using a custom perl script 'subsample\_fastq\_dir.pl' (Moraga 2018) to create datasets that had 90%, 80%, 70%, 60%, 50%, 40%, 30% of the data from the full dataset (average depth  $\approx$  12, 11, 10, 8, 7, 5 and 4 respectively).

As per Galla et al. (2019), the Tassel 5.0 GBSv2 pipeline (Glaubitz et al. 2014) was used to create a GBS tag database using the full kakī GBS dataset, specifying a k-mer length of 64, a minimum k-mer length of 20, a minimum Phred-score of 30 for the tag database plugin, and a minimum tag count of 10 for the tag export plugin. Bowtie2 (Langmead and Salzberg 2012) was used to align tags to the kakī reference genome. I ran the discovery SNP caller plugin with a minimum minor allele frequency of 0.05 and a minimum locus coverage of 0.1. The Tassel 5.0 GBSv2 production SNP caller plugin was then used to call SNPs for each of the subsampled kakī GBS datasets using the tag database generated from the full dataset.

The VCFs generated from the Tassel pipeline were then resampled to assess the effect of sample size. Simulated datasets of 1,000 replicates at each sample size ( $n = 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40$ ) were generated for each kakī GBS dataset using a custom perl script 'subsample\_vcf\_columns?.pl' (Moraga 2018) to randomly select a subset of individuals for each replicate.

Due to the lack of an available reference genome for Buller's albatross, the Stacks *de novo* pipeline (Catchen et al. 2011, Catchen et al. 2013) was used to genotype the entire dataset. I then used the 'populations' program in Stacks to split the dataset by population.

The VCFs for each populations dataset were then resampled to generate simulated datasets of 300 replicates at each sample size for the northern ( $n = 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28$ ) and the southern ( $n=2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 50, 60, 65$ ) populations. For each replicate, VCFtools v1.9 (Danecek et al. 2011) was used to generate random subsets of individuals.

### Calculating statistics for resampled datasets

A custom perl script 'calculate\_VCF\_stats.pl' (Moraga 2018) was then used to calculate total SNPs, total heterozygous SNPs and total heterozygous entries with a minimum allelic depth of 3 for all replicates in each simulated dataset. The dplyr package (Wickham et al. 2018) in the statistical software 'R' version 3.5.2 (R Core Team 2018) was used to collate these values and use them to calculate the proportion of heterozygous SNPs (total heterozygous SNPs / total SNPs) for each replicate. Scatterplots were then generated for each dataset to visually compare the distribution of these values for each sample size.

## Results

### Kakī resequencing

The total number of SNPs called for each kakī resequencing dataset increases with sample size in a non-linear way. The rate of increase in the number of SNPs with increasing sample size is similar for each of the datasets (Figure 2.1a-2.1c), however, fewer SNPs are called with lower depth (Figure 2.1d). The incremental increase in the total number of SNPs decreases with increasing sampling size, and each dataset asymptotes at approximately 8 samples regardless of depth.

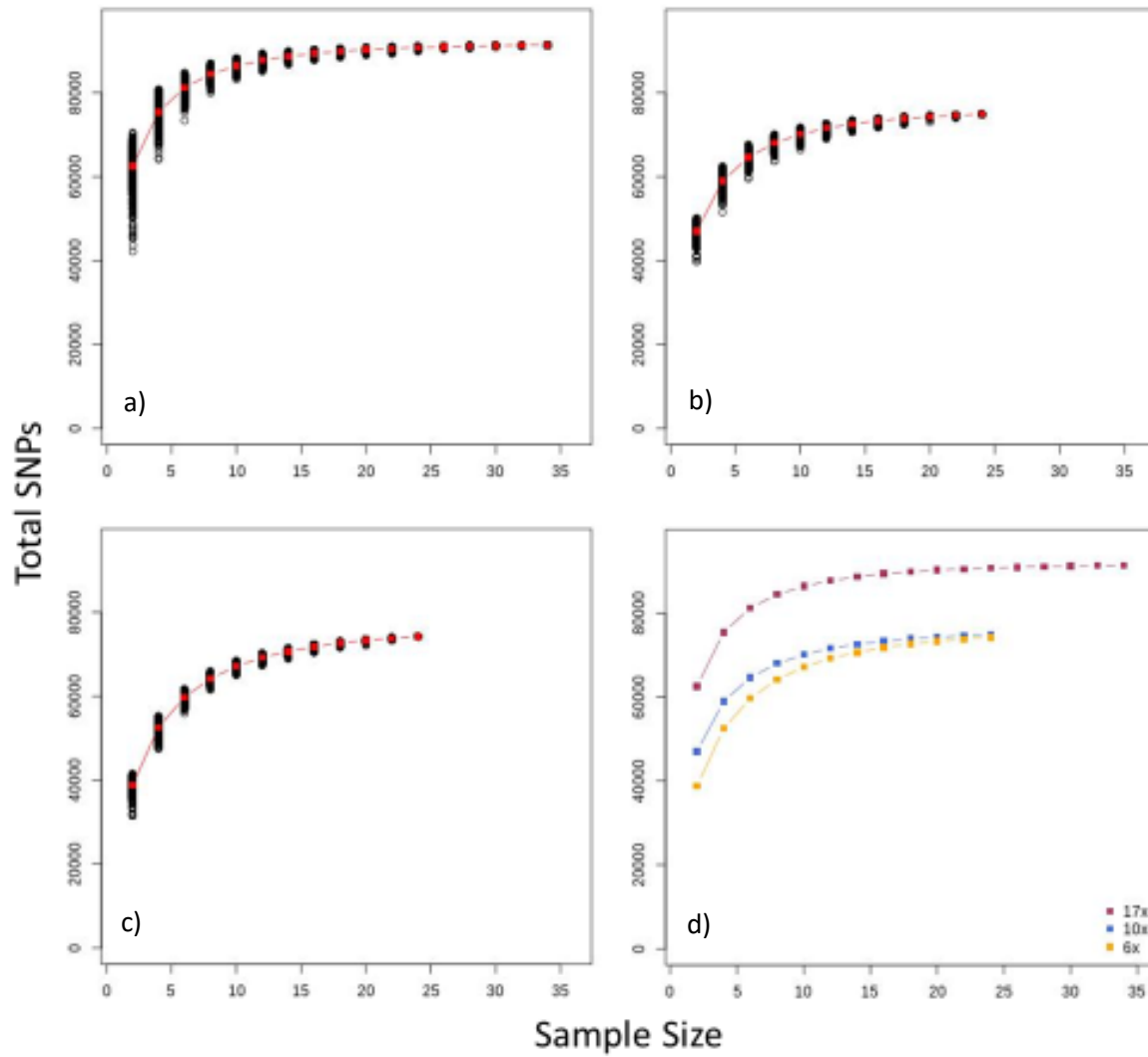


Figure 2.1: The impact of sample size and depth of coverage on the total number of SNPs with a minimum allelic depth of 3 within the kakī resequencing dataset. The lines are the mean total SNPs at each sample size (1,000 replicates) for: a) Full dataset ( $n = 34$ , mean depth = 17), b) High coverage samples ( $n = 24$ , mean depth = 10), c) High coverage samples ( $n = 24$ , mean depth = 6), d) Comparison of mean total SNPs from a), b) and c).

To get a better understanding of the effect of depth, I briefly investigated the impact it has on the number of heterozygous SNPs called per individual. As expected, reducing the depth of coverage also reduced the number of heterozygous SNPs called per individual (mean = 30424.97, 20246.913, 9165.399 for 17x, 10x and 6x datasets respectively; Figure 2.2).



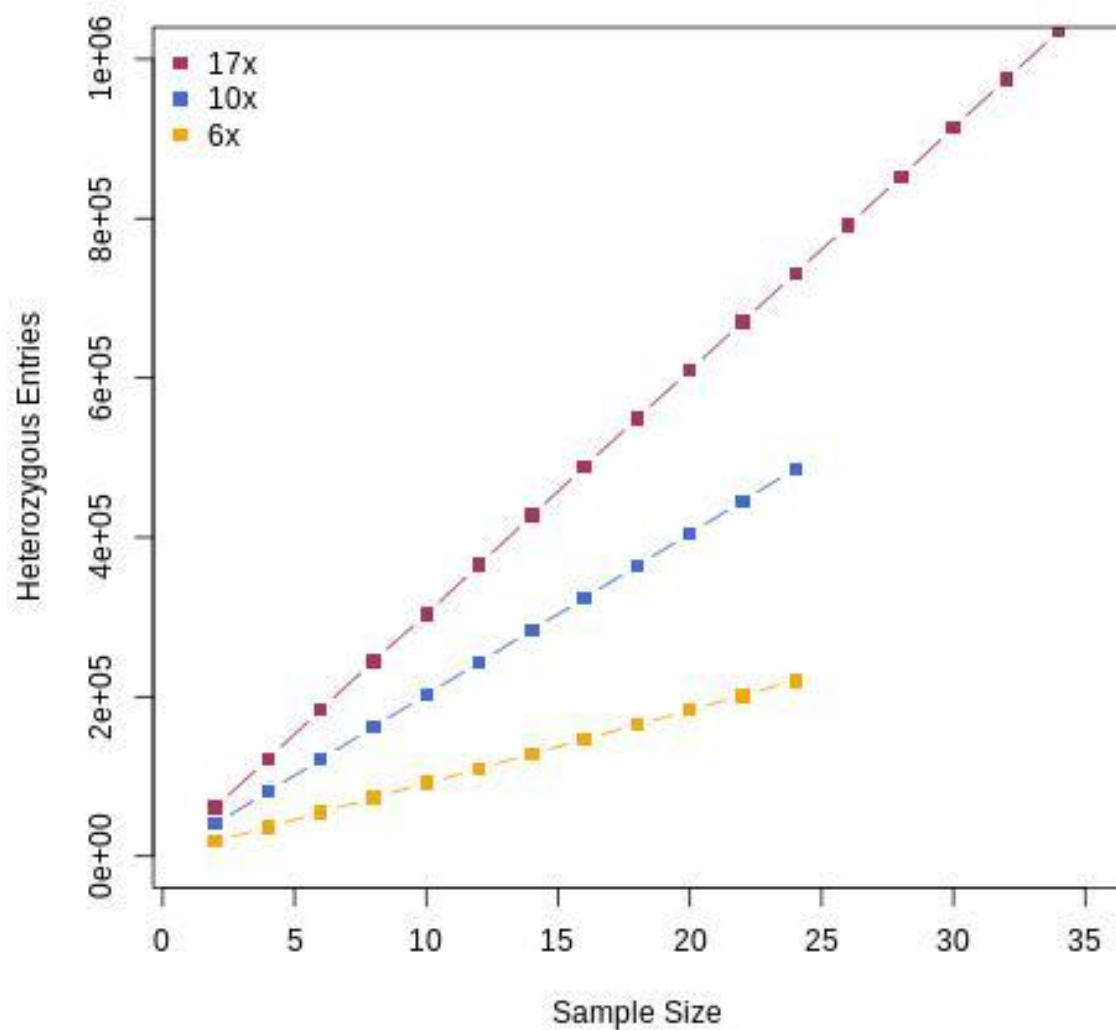


Figure 2.2: The impact of depth of coverage on the sum of all individual SNP calls that are heterozygous (i.e. heterozygous entries in the dataset). Each line corresponds to the mean values from each of the resampled kakī resequencing datasets and the slope of each line represents the mean number of heterozygous SNPs detected per individual.

Therefore, I assessed the effect of sample size on the proportion of SNPs that are heterozygous for each kakī resequencing resampled dataset. The proportion of heterozygous SNPs increases non-linearly with sampling size (Figure 2.3). However, when depth of coverage is too low, the rate of this

increase changes (Figure 2.3c). The higher depth datasets (17x and 10x) asymptote at approximately 8 samples whereas, in the lower depth dataset (6x), the sample size used ( $n=24$ ) was too small to reach an asymptote.

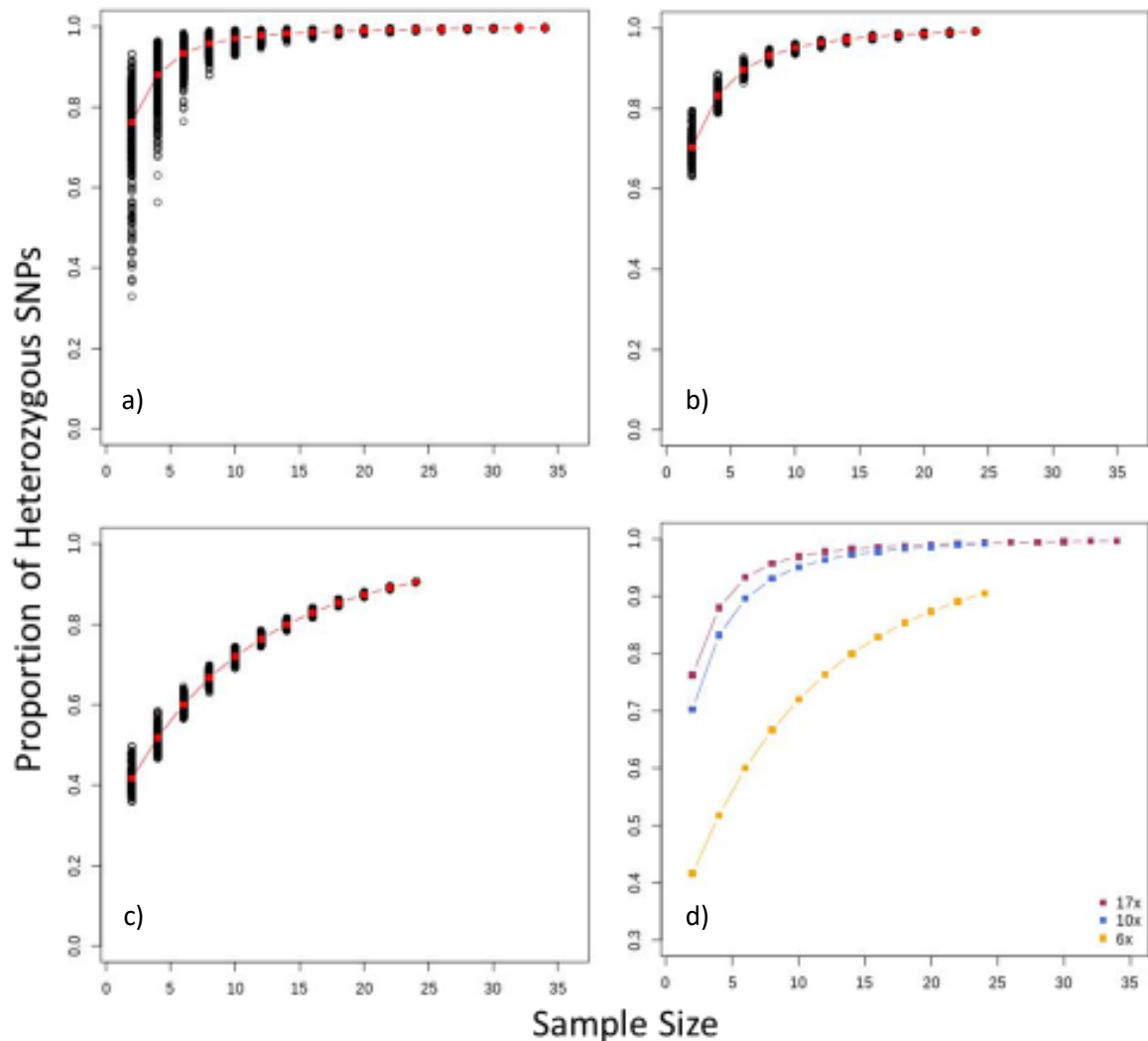


Figure 2.3: The impact of sample size and depth of coverage on the proportion of total SNPs that are heterozygous within the kakī resequencing dataset at a minimum allelic depth of 3 in both alleles.

The lines are the mean proportion of heterozygous SNPs at each sample size (1,000 replicates) for: a) Full dataset ( $n = 34$ , mean depth = 17), b) High coverage samples ( $n = 24$ , mean depth = 10), c) High coverage samples ( $n = 24$ , mean depth = 6), d) Comparison of mean proportion of heterozygous SNPs from a), b) and c). Note that the scale for d) has been reduced to focus on the relevant area of the figure.

## Kakī GBS

Here I assessed the effect of sample size on the proportion of SNPs that are heterozygous for kakī GBS resampled datasets at three different depth of coverage (14x, 10x and 7x). The proportion of heterozygous SNPs initially decreases non-linearly with sampling size (Figure 4). However, when depth of coverage is too low, the rate of this increase changes (Figure 3c). Each dataset (14x, 10x and 7x) asymptote at approximately 6, 8 and 16 samples respectively.

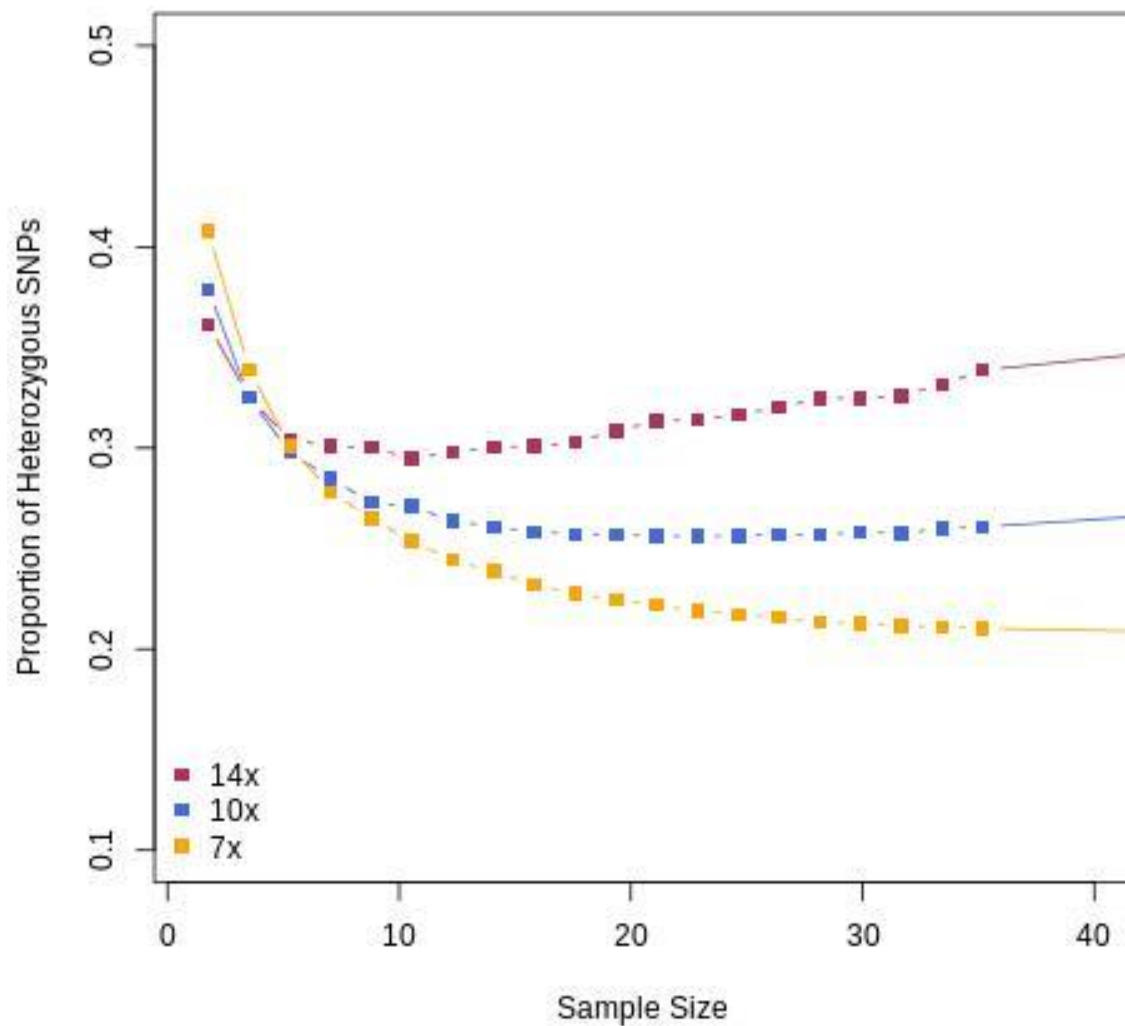


Figure 2.4: The impact of sample size and depth of coverage on the proportion of total SNPs that are heterozygous within the kakī GBS dataset with a minimum allelic depth of 3 in both alleles. The lines are the mean proportion of heterozygous SNPs at each sample size (1000 replicates) for each change in depth of coverage.

#### Buller's Albatross GBS

As with the other datasets, I once again assessed the effect of sample size on the proportion of heterozygous SNPs in the Buller's Albatross GBS resampled datasets. I found that the proportion of heterozygous SNPs increases non-linearly with sampling size for both the Northern and Southern populations (Figure 2.5). These datasets both asymptote at approximately 10-12 samples.

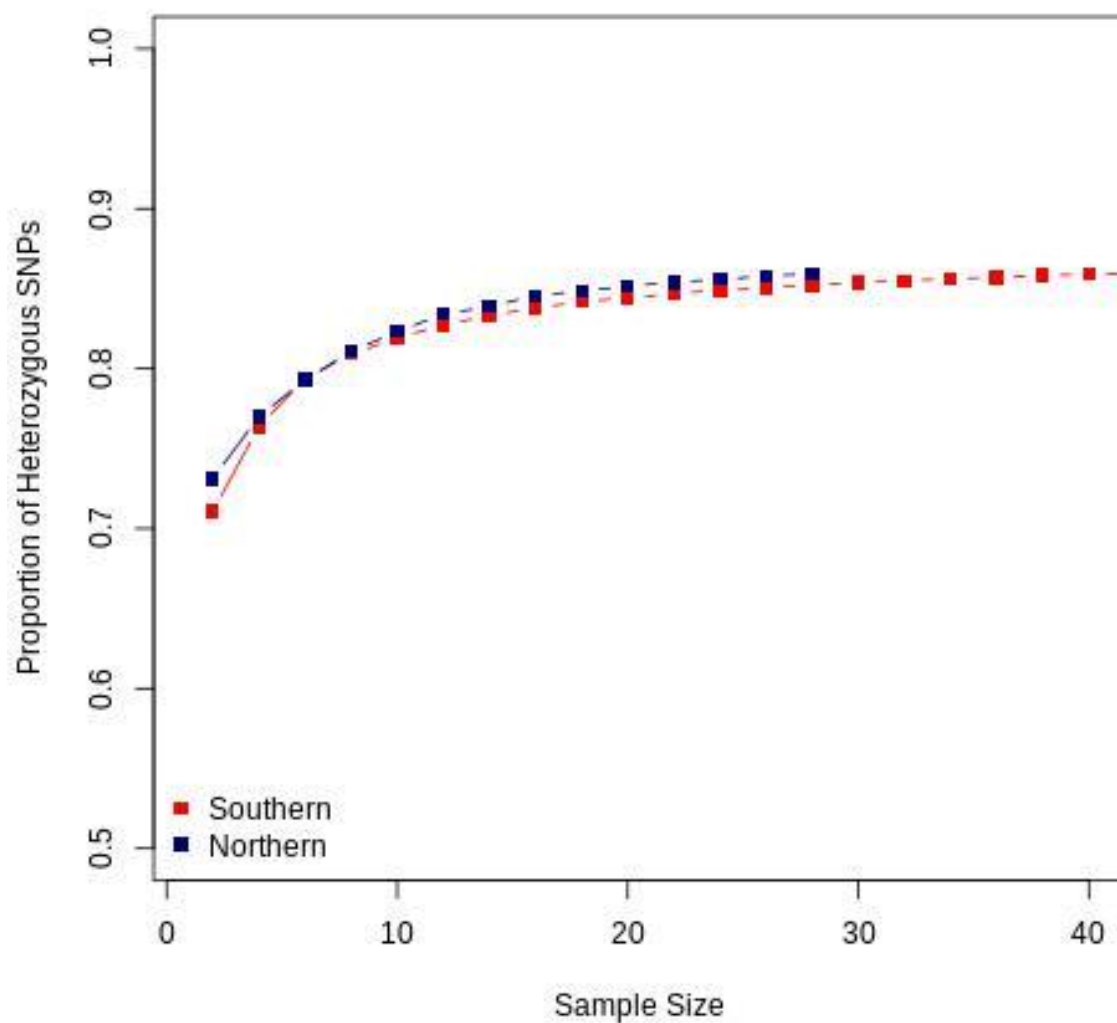


Figure 2.5: The impact of sample size on the proportion of total SNPs that are heterozygous with a minimum allelic depth of 3 in both alleles within GBS datasets from two populations of Buller's albatross. The lines are the mean proportion of heterozygous SNPs at each sample size (300 replicates) for each population.

## Discussion

Although depth of coverage reduced the total number of SNPs called in each of the resampled kakī resequencing datasets, the non-linear pattern of increasing total number of SNPs called with

increasing sample size was very similar for all resampled resequencing datasets (Figure 2.1). My intent for these analyses was to determine at which point is the increase in the number of SNPs obtained by sampling additional individuals outweighed by the increased cost of including these additional individuals. In these datasets that point is approximately 8 samples, regardless of the effect of depth, as increasing sample size beyond 8 yields relatively small increases in total SNPs. However, the total number of SNPs does not provide a measure of genomic diversity at the individual level, therefore, before making recommendations on sample size and depth of coverage, I investigated the effect that each of these have on heterozygosity (one of the most relevant measures of diversity for conservation genomic studies).

Not surprisingly, I found that depth of coverage has a significant effect on calling heterozygous SNPs (Figure 2.2), so the next step was to investigate the effect that increasing sample size has on the proportion of heterozygous SNPs detected. My results showed that for depths above 10x, the proportion of heterozygous SNPs increases with sample size in the same way as the total number of SNPs (Figure 2.3). Beyond 8 individuals sampled, additional samples have only a small effect on this measure of genome wide heterozygosity when depth of coverage is 10 or higher. However, when depth of coverage is low (i.e. 6x), a much larger sample size ( $n \geq 24$ ) would be necessary to accurately estimate genome wide heterozygosity.

The relationship between increasing sample size and proportion of heterozygous SNPs in the kakī GBS dataset is different to the other datasets. Here, the proportion of heterozygous SNPs initially decreases with increasing SNPs before reaching an asymptote and then slightly increasing (Figure 2.4). Here, it is appropriate to explore why this pattern exists only in the kakī GBS dataset and not in the kakī resequencing dataset or the Buller's Albatross GBS dataset. One potential explanation could be the way that SNPs were called for this dataset. A reference-guided approach in Tassel, which was used for the kakī GBS dataset, determines the potential physical positions of tags by mapping them against the reference genome, then tags positioned in the same physical location are aligned against

one another before identifying SNPs (Glaubitz et al. 2014). In contrast, a Stacks *de novo* approach, which was used for the Buller's albatross dataset, creates stacks from the raw data and aligns them to each other to create a catalog that is used as a reference for calling SNPs (Catchen et al. 2011). In this sense, the *de novo* approach used for Buller's Albatross dataset acts more like the reference-guided approach used with WGR kakī dataset. Regardless of the explanation, the inference that can be drawn from the kakī GBS is concordant with that of the kakī resequencing data: when depth of coverage is 10 or higher, increasing sample size beyond 8 individuals had negligible effects on the proportion of heterozygous SNPs (Figure 4).

The results for the northern and southern Buller's Albatross GBS datasets were similar to the WGS and GBS kakī datasets, with the proportion of heterozygous SNPs increasing with sample size until reaching a point where the curve asymptotes (Figure 2.5). However, this point is at a larger sample size ( $n = 10-12$ ) than in both kakī datasets ( $n = 8$ ), which provides preliminary support for the hypothesis that a more diverse populations require larger sample sizes.

## Conclusions

When using either a reduced-representation approach like GBS or a whole genome resequencing approach for critically endangered species such as kakī or kōwaro, I recommend sampling a minimum of 8 individuals per population, with a minimum depth of coverage of 10. Based on the preliminary data for Buller's albatross, it would be premature to make recommendation for non-threatened species, but it is reasonable to suggest that more individuals per population will need to be sampled. Regarding more specific recommendations for kōwaro, because the number of samples required per populations is relatively low (compare  $n = 25-30$  for microsatellites to  $n = 8$  for SNPs), an additional cost consideration is the trade-off between number of populations samples and depth of coverage per individual. For example, the difference between sampling more populations where individuals samples are samples sequenced at lower depth, or sampling fewer populations where

individuals are sequenced at higher depth: the cost of WGR kōwaro at 20x is \$450 per sample, compared to \$315 per sample for WGR at 15x so sequencing all individuals at lower depth may warrant the inclusion of additional populations, especially if the conservation genomic questions of interest is population genomic structure.

These combined results are consistent with other studies that recommend relatively low sample sizes per population when using genomic data (e.g. Willing et al. 2012; Gaughran et al. 2017; Nazareno et al. 2017) and this research is the first to empirically estimate adequate sample sizes for population studies in critically endangered species using whole genome resequencing data.

## References

- Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. and Postlethwait, J.H., 2011. Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, genomes, genetics*, 1(3), pp.171-182.
- Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A. and Cresko, W.A., 2013. Stacks: an analysis tool set for population genomics. *Molecular ecology*, 22(11), pp.3124-3140.
- Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. and McVean, G., 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15), pp.2156-2158.
- Flesch, E.P., Rotella, J.J., Thomson, J.M., Graves, T.A. and Garrott, R.A., 2018. Evaluating sample size to estimate genetic management metrics in the genomics era. *Molecular ecology resources*, 18(5), pp.1077-1091.



Fuentes-Pardo, A.P. and Ruzzante, D.E., 2017. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular ecology*, 26(20), pp.5369-5406.

Galla, S.J., Buckley, T.R., Elshire, R., Hale, M.L., Knapp, M., McCallum, J., Moraga, R., Santure, A.W., Wilcox, P. and Steeves, T.E., 2016. Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances. *Molecular ecology*, 25(21), pp.5267-5281.

Galla, S.J., Forsdick, N.J., Brown, L., Hoepfner, M., Knapp, M., Maloney, R.F., Moraga, R., Santure, A.W. and Steeves, T.E., 2019. Reference Genomes from Distantly Related Species Can Be Used for Discovery of Single Nucleotide Polymorphisms to Inform Conservation Management. *Genes*, 10(1), p.9.

Gaughran, S.J., Quinzin, M.C., Miller, J.M., Garrick, R.C., Edwards, D.L., Russello, M.A., Poulakakis, N., Ciofi, C., Beheregaray, L.B. and Caccone, A., 2018. Theory, practice, and conservation in the age of genomics: The Galápagos giant tortoise as a case study. *Evolutionary applications*, 11(7), pp.1084-1093.

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q. and Buckler, E.S., 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PloS one*, 9(2), p.e90346.

Hagen, E.N., Hale, M.L., Maloney, R.F. and Steeves, T.E., 2011. Conservation genetic management of a critically endangered New Zealand endemic bird: minimizing inbreeding in the Black Stilt *Himantopus novaezelandiae*. *Ibis*, 153(3), pp.556-561.

Hale, M.L., Burg, T.M. and Steeves, T.E., 2012. Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PloS one*, 7(9), p.e45170.

Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), p.357.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), pp.2078-2079.

Miyamoto, N., Fernández-Manjarrés, J.F., Morand-Prieur, M.E., Bertolino, P. and Frascaria-Lacoste, N., 2008. What sampling is needed for reliable estimations of genetic diversity in *Fraxinus excelsior* L.(Oleaceae)?. *Annals of forest science*, 65(4), p.1.

Moraga, R. SubSampler\_SNPcaller Available online:

[https://github.com/Lanilen/SubSampler\\_SNPcaller](https://github.com/Lanilen/SubSampler_SNPcaller)

Nazareno, A.G., Bemmels, J.B., Dick, C.W. and Lohmann, L.G., 2017. Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Molecular Ecology Resources*, 17(6), pp.1136-1147.

Pruett, C.L. and Winker, K., 2008. The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. *Journal of Avian Biology*, 39(2), pp.252-256.

R Core Team 2018. R: A language and environment for statistical computing. Retrieved from

<https://www.R-project.org/>

Robertson, H.A., Baird, K., Dowding, J.E., Elliott, G.P., Hitchmough, R.A., Miskelly, C.M., McArthur, N., O'Donnell, C.F.J., Sagar, P.M., Scofield, R.P., Taylor, G.A., 2017. *Conservation status of New Zealand birds, 2016. New Zealand Threat Classification Series 19*. Department of Conservation, Wellington. 23 p.

Ryman, N., Palm, S., André, C., Carvalho, G.R., Dahlgren, T.G., Jorde, P.E., Laikre, L., Larsson, L.C., Palmé, A. and Ruzzante, D.E., 2006. Power for detecting genetic divergence: differences between statistical methods and marker loci. *Molecular Ecology*, 15(8), pp.2031-2045.

Willing, E.M., Dreyer, C. and Van Oosterhout, C., 2012. Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PloS one*, 7(8), p.e42649.

Wickham, H., François, R., Henry, L., Müller, K. 2018. Dplyr: A Grammar of Data Manipulation. R package version 0.7.6. Retrieved from <https://CRAN.R-project.org/package=dplyr>

Yan, L. and Zhang, D., 2004. Effects of sample size on various genetic diversity measures in population genetic study with microsatellite DNA markers. *Dong wu xue bao.[Acta zoologica Sinica]*, 50(2), pp.279-290.

## Chapter Three: Embedding kaupapa Māori principles in genomic research of taonga species: a conservation genomics case study

Levi Collier-Robinson (Ngāi Tūāhuriri, Poutini Ngāi Tahu)<sup>1</sup>, Aisling Rayne<sup>1</sup>, Makarini Rupene (Ngāi Tūāhuriri, Ngāi Tahu)<sup>2</sup>, Channell Thoms (Ngāti Kurī, Ngāi Tahu)<sup>1</sup>, Tammy Steeves<sup>1</sup>

<sup>1</sup> School of Biological Sciences, University of Canterbury, Christchurch, New Zealand;

<sup>2</sup> Ngāi Tahu Research Centre, Christchurch, New Zealand

### Abstract

In Te Ao Māori, genomic data obtained from taonga species have whakapapa and are therefore taonga in their own right. Thus, genomic data are tapu and best studied using kaupapa Māori principles. We contend it is the responsibility of researchers working with genomic data from taonga species to move beyond one-off Māori consultation toward building meaningful relationships with relevant Māori communities. Here, we reflect on our experience embedding kaupapa Māori principles in genomics research as leaders of a Biological Heritage National Science Challenge project entitled “Characterising adaptive variation in Aotearoa New Zealand’s terrestrial and freshwater biota”. We are co-developing a culturally-responsive evidence-based position statement regarding the benefits and risks of prioritising adaptive potential to build resilience in threatened taonga species, including mahinga kai species destined for customary or commercial harvest. To achieve this, we co-developed a research programme with Ngāi Tūāhuriri that integrates Mātauranga Māori with emerging genomic technologies and extensive ecological data for two taonga species, kōwaro

(*Neochanna burrowsius*) and kēkēwai (*Paranephrops zealandicus*). The foundation of our research programme is an iterative decision-making framework that includes tissue sampling as well as data generation, storage and access. Beyond upholding the promises made in Te Tiriti o Waitangi, we contend the integration of kaupapa Māori principles in genomics research will enhance the recovery of taonga species and enable the realisation of Māori values.

*Keywords:* Mātauranga Māori, indigenous knowledge, whakapapa, rangatiratanga, tohungatanga, whanaungatanga, kaitiakitanga, decision-making framework, kōwaro, kēkēwai

*Lay summary:* To provide an example of an effective approach for building meaningful relationships with relevant Māori communities for mutual benefit, we reflect on our experience embedding kaupapa Māori principles in our research on the whakapapa of two taonga species and provide an iterative decision-making framework that is broadly applicable to genomic research on taonga species in Aotearoa New Zealand.

Te Tiriti o Waitangi (1840) is a crucial founding document that frames the relationship between Māori and the Crown in Aotearoa New Zealand. Thus, Te Tiriti o Waitangi is at the forefront of all interactions between Māori and Pākehā. Article Two of Te Tiriti o Waitangi guarantees to Māori the rangatiratanga over their taonga and ensures that the rights of both Māori as tangata whenua and Pākehā are preserved. Historically there have been numerous actions from the Crown that breached these promises of Te Tiriti o Waitangi (Walker 1990). Iwi Māori fought for generations to settle these historical grievances which led to the Treaty of Waitangi Act 1975 and the establishment of the Waitangi Tribunal (Walker 1990). Now, many iwi are moving beyond settling their historical grievances into an era of growth and partnership. For example, in his address at the Ngāi Tahu Treaty Commemoration Hui at Ōnuku Marae (2019), Tā Tipene O'Regan stated:

*“...we have now reached a point where we must see ourselves no longer as the damaged and dispossessed victims of the New Zealand Project but as part of, and contributors to, the development of what this nation might yet become.”*

As a living document in Aotearoa, Te Tiriti o Waitangi has led to government policies and Waitangi Tribunal Reports that provide a clear mandate for research partnership. Of particular relevance, Vision Mātauranga (Ministry of Research, Science and Technology 2007) seeks to ‘unlock the science and innovation potential of Māori knowledge, people and resources’ and Ko Aotearoa Tēnei/This is New Zealand, a report into the WAI 262 claim conventionally known as WAI 262 (<http://www.waitangitribunal.govt.nz/>), extends the scope of Te Tiriti o Waitangi to claim the rights of Māori to ngā taonga katoa (reviewed in Ataria et al. 2018). In Te Ao Māori, ngā taonga katoa refers to all things that are treasured by Māori, including indigenous culture, knowledge, flora and fauna. Thus, Te Tiriti o Waitangi is an important consideration for all research conducted in Aotearoa, especially research involving taonga species.

As researchers based at The University of Canterbury, we fall within the territory of Ngāi Tahu who are mana whenua for most of the South Island. Ngāi Tūāhuriri is the hapū that are mana whenua from Hurunui to Hakatere and inland to the Main Divide. Te Rūnanga o Ngāi Tahu were able to negotiate treaty settlements with the Crown earlier than most iwi. Since then, they have experienced significant growth and development. However, it is important to recognise that not all tribal groups have had the same experiences, and each iwi and hapū are at a unique stage of development. These factors can determine the capacity for mana whenua to be involved in taonga species research, but it does not affect the relevance of the research to them. Furthermore, for researchers, developing a deeper understanding of the needs, aspirations and circumstances of relevant iwi or hapū enables them to better apply their skills to research questions that are of interest to mana whenua.

The following quote from Kemps Deed, the largest Ngāi Tahu land purchase by the Crown details the importance of mahinga kai to Ngāi Tahu:

*“Ko ō mātou kāinga nohoanga, ko ā mātou mahinga kai, me waiho mārie  
mō ā mātou tamariki, mō muri iho i a mātou.”*

*“Our places of residence, cultivations and food gathering places must still  
be left to us, for ourselves and our children after us”.*

As a reminder of past breaches of Te Tiriti o Waitangi and a forecast of the future direction for the iwi, it led to the following quote which now acts as the guiding whakataukī for Ngāi Tahu:

*“Mō tatou, ā, mō kā uri ā muri ake nei”*

*“For us, and our descendants after us”*

Kaupapa Māori research is based on several key principles and philosophies that are applicable to all research conducted in Aotearoa New Zealand. It is an approach that has arisen from Te Tiriti o Waitangi that enables researchers to consider ethical, methodological and cultural issues from another perspective throughout the research process (Pihama et al. 2002; Smith 1997; Smith 2013; Walker et al. 2006). Kaupapa Māori research originated within an education context (Smith 1997) and has since been expanded by several Māori theorists to encompass research in a more general sense (Pihama 2012; Pihama et al. 2002; Smith 2013). Although there are many interconnected kaupapa Māori research principles, some may be more relevant than others within any given context.

Ngāi Tahu and Ngāi Tūāhuriri place a strong emphasis on embodying the following core values: whakapapa, whanaungatanga, manaakitanga, tikanga, tohungatanga, rangatiratanga and kaitiakitanga. All of these are either kaupapa Māori principles themselves or encompassed by them. Below, we use Te Ao Māori to frame these core values and to highlight four key kaupapa Māori principles applicable to genomic research involving taonga species with a particular focus on Ngāi Tahu interests.

## Ngā taonga tuku iho

This context provided by Article Two of Te Tiriti o Waitangi is about acknowledging the validity and relevance of Māori ways of knowing and understanding the world (Pihama et al. 2002). There are several interconnected concepts in Te Ao Māori that researchers should consider when working with taonga species that may lead to opportunities to integrate Mātauranga Māori and western science.

Te Reo Māori is an excellent starting point. Te Ao Māori is entrenched in the language, including Māori place names, whakataukī, and associated stories (Wehi et al. 2009; Whaanga et al. 2018). In



contrast to the analytical nature of the English language, Te Reo Māori is filled with symbolism and emotional embellishment that allows Māori to intuitively grasp complex concepts. Embracing the strengths of both languages can lead to co-development of research frameworks relevant to both Māori and non-Māori (Mercier 2018; Walker et al. 2006). For example, mauri is the life force found in all things: it is the essential quality and vitality of an entity, whether that is a physical object, an individual or an ecosystem (Hikuroa et al. 2011). The integration of Mātauranga Māori and western science can enable frameworks that seek to maintain and enhance mauri and other Māori values (Harmsworth and Tipa 2006; Hikuroa et al. 2011; Hudson et al. 2016; Rainforth and Harmsworth 2019).

Tikanga Māori is about the appropriate way to operate within a Māori context; including customary practices, protocols and ethics (Mead 2003). It dictates how Māori interact with each other, and with their environment. Tapu and noa are Māori concepts that fundamentally shape tikanga Māori. They are complex and multifaceted, but uncomplicated. Tapu refers to that which is sacred, special, forbidden or restricted; whereas noa is the inverse of tapu and refers to the common and unrestricted (Mead 2003). All taonga are inherently tapu, and tikanga therefore determine how people interact with our taonga.

Mātauranga Māori is traditionally passed down orally through pūrākau, waiata, pepeha and whakataukī, or visually through mahi toi (Hikuroa 2017). These ancestral stories are then contextualised using whakapapa (Tau 2001). Although many pūrākau are myths and heavily symbolic in nature, they still serve the practical function of passing on Māori culture and the knowledge of the natural world through a Māori world view (Hikuroa 2017). They also explain the relationship that tangata whenua share with the world around them by associating their tupuna with specific aspects of the environment.

For researchers with a genuine interest in embedding Mātauranga Māori in their research, developing a general understanding of Te Ao Māori is invaluable. Moreover, we argue it is imperative for researchers to be mindful of local context, particularly when working with the whakapapa of taonga species.

Whakapapa is generally defined as genealogy, but in Te Ao Māori, it encompasses much more than that (Te Rito 2007). It layers the contemporary, historical, spiritual and mythological aspects of heritage (Tau 2001). Whakapapa is critical in shaping how Māori view the world, and from a traditional Māori perspective, all life on Earth can be traced back through whakapapa (Tau 2001; Te Rito 2007).

Although the most common application of whakapapa in a modern context is to describe family pedigrees, whakapapa is not limited to people. The whakapapa of people, animals and plants; mountains, rivers and winds are all interconnected and explain these complex relationships through a Māori lens (Tau 2001). There are a multitude of similarities between whakapapa and a range of western science disciplines, the most literal being DNA-based research.

DNA is a physical expression of whakapapa. Like DNA, whakapapa is unique to any one hierarchical group. This uniqueness inherently renders whakapapa - and by extension, DNA - as a taonga and something that is tapu (Beaton et al. 2017; Hudson et al. 2016). Therefore, tikanga should influence the way that genetic and genomic data are generated and used. However, not all traditional tikanga practices apply to something so novel. Indeed, as modern western science continues to develop new methods, the tikanga surrounding it will also change. Thus, there is a need for Māori communities to be involved with emerging DNA technologies so actions appropriate for Aotearoa can be co-developed by researchers and tangata whenua.

The whakapapa of Māori deities can be viewed as a hierarchical classification of the origin of both the abiotic and biotic aspects of the environment. There are similarities in these ancient creation

stories across iwi, but subtle differences between them reflect the need for Māori to describe novel landscapes in new ways. Whakapapa in these settings is used as a tool to enrich Mātauranga Māori within local contexts. For example, the story of Ranginui, Papatūānuku and their children is a very common Māori creation narrative (Reed 2004). However, Pokoharuatēpō, the first wife of Ranginui and the mother of Aoraki have a special significance to Ngāi Tahu. In this narrative, the creation of what is now known as Te Waipounamu or the South Island is attributed to the wreckage of Te Waka o Aoraki when he and his brothers journeyed to meet their new step-mother Papatūānuku. Aoraki and his brothers eventually turned to stone on top of their overturned canoe where they now form the principal peaks of the Southern Alps. This perspective of the South Island landscape is unique to Ngāi Tahu and this whakapapa illustrates the importance of Aoraki (Mt Cook) to the people of Ngāi Tahu. By extension, researchers working in the Ngāi Tahu takiwā need to be mindful of the local narrative, for example, by developing an understanding of the significance of place names and the stories behind them (e.g., Kā Huru Manu, <http://www.kahurumanu.co.nz/>).

### Key kaupapa Māori principles for genomic research on taonga species

A major focus of kaupapa Māori research is enabling **rangatiratanga** by providing tangata whenua with the autonomy and authority to practice and share their own culture, knowledge and other taonga in their own way (Pihama et al. 2002; Smith 1997). Within a research context, it enables Māori to shape how their taonga are researched.

*“He aha te mea nui o te Ao? He tangata, he tangata, he tangata.”*

*“What is the most important thing in the world? It is the people, it is the people, it is the people.”*

**Whanaungatanga** represent our relationships with one another and enables kaupapa Māori research through the process of building and maintaining meaningful partnerships with tangata whenua that are necessary for collaborative projects and an expression of rangatiratanga (Smith 2013; Walker et al. 2006). It lies at the core of Māori culture and society, therefore, whakawhanaungatanga is the most important step for researchers looking to engage with Māori in a meaningful way. Although there are frameworks available to assist researchers (e.g. Hudson and Russell 2009; Smith 2013), building significant relationships with Māori cannot be reduced to simple step-by-step procedures. However, these frameworks can help researchers to recognise and acknowledge the unique culture and tikanga of each iwi, hapū and whānau that are involved in the research.

**Kaitiakitanga** is a term that has become widely used in mainstream New Zealand regarding species conservation and ecosystem restoration. However, it encompasses more than just conserving species or restoring ecosystems: kaitiakitanga includes everything that is taonga to tangata whenua, including knowledge, culture and language (Lyver and Tylianakis 2017, Wehi and Lord 2017, Wehi et al. 2018, Lyver et al. 2019). Research focused on recovering taonga species, particularly mahinga kai species, has the potential to enhance these interconnected elements. Kaitiakitanga of mahinga kai includes the environment, language, culture and knowledge associated with harvesting practices. Thus, research that aims to enhance species recovery can facilitate the revitalisation of the language and practices associated with these species.

Tohunga were traditionally expert practitioners in a given field that gave direction to others and helped to develop others. Therefore, **tohungatanga** encourages whānau to develop capability and capacity while supporting the development of others. The very nature of science collaboration with mana whenua achieves tohungatanga, as it builds expertise within iwi and hapū to pursue knowledge and ideas that will enable them to strengthen and grow. Furthermore, whanaungatanga is realised through genuine co-development of research ideas and active engagement throughout

research process. In doing so, rangatiratanga and kaitiakitanga are also realised because the authority and sovereignty that mana whenua have over their own taonga are recognised.

As researchers with pre-existing relationships with Ngāi Tahu and Ngāi Tūāhuriri, we were given the opportunity to incorporate these key kaupapa Māori principles in a new scope of work involving genomic research of threatened taonga species, and together with mana whenua frame a narrative that speaks to the subtleties of Te Ao Māori often overlooked by typical western science practice. Here, we share this narrative, not as a template to be followed or as a series of boxes to be ticked, but as an example of one way to better enhance the recovery of taonga species.

## Genomic research

Genetics and genomics approaches for studying DNA have become invaluable tools for many biological disciplines, including the conservation of threatened species (reviewed in Galla et al. 2016). New technologies are rapidly expanding our ability to extract, generate and understand DNA. As these technologies become more efficient, they become more affordable and accessible too. Here, we provide an overview of genetics and genomics, and outline several necessary considerations when generating these data from taonga species.

Traditionally, conservation genetic studies use a small set of genetic markers scattered throughout the genome to estimate genetic diversity within and between populations in an effort to inform conservation management (Frankham et al. 2010). These strategies are generally implemented in a way that seeks to reduce adverse effects associated with small, isolated populations by minimising inbreeding and the loss of genetic diversity (Frankham et al. 2017). However, there are limitations to using only a small number of genetic markers within a genome that has millions, if not billions, of DNA base pairs, including variation at a small number of selectively neutral markers unlikely being

representative of genome-wide variation and, at best, only being able to be used as a proxy for the ability of a species to adapt to changing environments (Allendorf et al. 2010; Ouborg et al. 2010; Funk et al. 2012; Defaveri et al. 2013).

High-throughput DNA sequencing is rapidly changing the way that we address conservation genetic questions. These new technologies are enabling the generation of reference genomes, as well as the characterisation of many thousands of single nucleotide polymorphisms (SNPs), for non-model species (e.g., Galla et al. 2019). The ability to generate a large number of genome-wide markers within and among natural populations is enabling researchers to address old questions at higher resolution (e.g., estimating relatedness; Lemopoulos et al. 2019) and to tackle entirely new ones (e.g., characterising adaptive potential; Chen 2019; de Villemereuil et al. 2019).

Regardless of whether researchers generate handfuls of microsatellites versus thousands of SNPs, or single reference genomes versus numerous re-sequenced genomes, the status of these data as taonga remains the same. As a result, data security and management of genetic and genomic data from taonga species has become paramount and considered discussions from a Māori perspective are underway across Aotearoa (e.g., SING – Aotearoa, <https://www.singaotearoa.nz/>). These include discussions that will lead to the development of guidelines for genomic research from taonga species led by Genomics Aotearoa (Te Nohonga Kaitiaki, <https://www.genomics-aotearoa.org.nz/projects/te-nohonga-kaitiaki>). In the meantime, there are growing initiatives in Aotearoa that seek to manage access and storage of genomic data from taonga species with appropriate kaitiakitanga (Catanach et al. 2019, Galla et al. 2019, Wellenreuther et al. 2019; also see: <https://www.genomics-aotearoa.org.nz/data> and <http://www.uconsert.org/data>).

## Case study

As leaders of a Biological Heritage National Science Challenge project entitled “Characterising adaptive variation in Aotearoa New Zealand’s terrestrial and freshwater biota”, we co-developed a research programme with mana whenua that is integrating Mātauranga Māori with emerging genomic technologies and extensive ecological data to characterise adaptive potential - or the ability to adapt to environmental change - in two taonga species, kōwaro (*Neochanna burrowsius*) and kēkēwai (*Paranephrops zealandicus*). We are combining these data with three additional focal species to co-develop a culturally-responsive, evidence-based position statement regarding the benefits and risks of prioritising adaptive potential to build resilience in threatened taonga species, including mahinga kai species destined for customary or commercial harvest. The foundation of our research programme is an iterative decision-making framework that embeds kaupapa Māori relevant principles. It begins by framing the research narrative in partnership with mana whenua followed by active engagement to make decisions regarding tissue sampling as well as data generation, storage and access, and ends by sharing the research narrative in partnership with mana whenua (Figure 3.1). Below, we show how we applied the iterative decision-making framework to our conservation genomic research on kōwaro and kēkēwai. We also demonstrate how this framework is broadly applicable to all genomic research on taonga species.



Figure 3.1. An iterative decision-making framework, indicating relevant kaupapa Māori principles and focal areas for active engagement with mana whenua regarding genomic research on two threatened taonga species, kōwaro (*Neochanna burrowsius*) and kēkēwai (*Paranephrops zealandicus*). See text for details.

The first taonga species that we co-identified with Ngāi Tūāhuriri is kōwaro (Canterbury mudfish; *Neochanna burrowsius*), one of the most endangered endemic freshwater fish species in Aotearoa, currently classified as “Nationally Critical” by the Department of Conservation (Dunn et al. 2018). Kōwaro are restricted to the Canterbury plains, and they have a fragmented distribution between the Rakahuri (Ashley) and Waitaki river catchments (Cadwallader 1975; O’Brien and Dunn 2007). Range restriction and severe loss of habitat due to land use intensification in Canterbury are key factors contributing to its current conservation status (Barrier 2003; Dunn et al. 2018; O’Brien and Dunn 2007). The continued threat of local extirpation across its range has led to a call for urgent conservation action (Dunn et al. 2018).



One such conservation action is a translocation project based at Tūhaitara Coastal Park. The park was established by Te Kōhaka o Tūhaitara Trust following the Ngāi Tahu settlement with the crown and it encompasses Te Tiriti o Waitangi; a collaborative effort between the people of the treaty. The area is rich in Ngāi Tūāhuriri history and mahinga kai, and kōwaro are an integral part of this ecosystem. Kōwaro was co-selected for our project because a conservation genomics approach is likely to enhance conservation outcomes to help preserve kōwaro as part of the unique biodiversity of Tūhaitara Coastal Park.

Endemic to Aotearoa, kēkēwai (freshwater crayfish / kōura; *Paranephrops zealandicus*) are a declining taonga species found in lakes, streams and ponds in the east and south side of Te Wai Pounamu / South Island as well as Rakiura / Stewart Island (Grainger et al. 2018). The *Paranephrops* genus has been a traditional food source for Māori across Aotearoa for centuries and has more recently been the focus of aquaculture initiatives for customary and commercial harvest (Parkyn and Kusabs 2007; Monk 2017).

Although kēkēwai as a species is not at immediate risk of extinction, land use intensification in Canterbury is fragmenting kēkēwai populations and driving local decline (Thoms 2016). Most remaining populations within the Ngāi Tūāhuriri takiwā now face extirpation. In addition to informing the recovery of declining wild populations, kēkēwai was co-selected for our project because a conservation genomics approach can enhance customary and commercial harvest, making these practices more sustainable so that they can continue for generations to come (Kristensen et al. 2015; Galla et al. 2016).

After framing the research narrative for each species, we discussed sampling design with Ngāi Tūāhuriri, including tissue sampling at sites of cultural significance traditionally used for mahinga kai. Doing so is especially important when generating reference genomes because these invaluable resources are a physical representation of Ngāi Tūāhuriri whakapapa. For the kōwaro reference

genome, the obvious choice of location was within Tūhaitara Coastal Park. However, due to the uncertain status of this small, fragmented and isolated population, we collectively decided to lethally sample a single individual from a larger, healthier population elsewhere in the Ngāi Tūāhuriri takiwā. For kēkēwai, we lethally sampled two individuals approximately one year apart from a small stream near Tuahiwi at the heart of the Ngāi Tūāhuriri takiwā.

Sampling animals has its own tikanga and practices within western science, typically regulated by animal ethics committees. Māori have their own tikanga and Mātauranga for taonga species and have harvesting practices that are excellent for sampling (Kusabs and Quinn 2009). As a mahinga kai species, kēkēwai allowed us to integrate Mātauranga Māori into a modern context to sample effectively and ethically. We used bundled bracken ferns to create tau kōura as a traditional method of harvest to capture kēkēwai (Parkyn and Kusabs 2007; Kusabs and Quinn 2009) and the maramataka (Māori lunar calendar) to determine favourable days for collection.

The tissue sampled from kōwaro and kēkēwai has value in the information it contains, therefore the tissue itself is taonga (Hudson et al. 2016). Ngāi Tūāhuriri have the rangatiratanga to determine the tikanga for generating the reference genomes for these species. As researchers with the relevant expertise, it was our responsibility to clearly communicate the benefits and risks of any given approach. Thus far, we have focused on whether to generate the reference genomes here in Aotearoa or overseas. After considering data quantity, data quality, data security, turnaround time and cost, we made the collective decision to send DNA for both kōwaro and kēkēwai to a trusted provider overseas with extensive experience handling culturally sensitive material. By including mana whenua in this way, we promote rangatiratanga while building tohungatanga around the research. In addition to generating genomic data, we are characterising the ecological characteristics of kōwaro and kēkēwai habitats. It is important to note that this ecological data also has its own mauri. It adds another layer to the whakapapa and should therefore be treated with the same manaakitanga (e.g., Bond et al. 2019).

During our research we have encountered existing or new transcriptome data that can be used to supplement the reference genomes for both kōwaro and kēkēwai (P. Dearden unpublished data, Wallis and Wallis 2014). Prior to the inclusion of these data, which are also taonga, we are engaging with relevant mana whenua. Related to this, as we expand our research beyond the Ngāi Tūāhuriri takiwā, we are acknowledging the interests and aspirations of other hapū and papatipu rūnanga within Ngāi Tahu and beyond, while being mindful that whakawhanaungatanga will be experienced uniquely with each different group.

Te Tiriti o Waitangi promises that tangata whenua retain the rangatiratanga over their own taonga which includes the whakapapa of taonga species. Genetic data have traditionally been shared openly on globally accessible databases. Rapid advancements in the field of genomics has led to data that are more complex and valuable. Therefore, rangatiratanga has become increasingly important in how knowledge and data from taonga species are shared. The challenge of upholding Te Tiriti o Waitangi is a national one, but it is tangata whenua who ultimately have the right to determine how their own whakapapa is shared. As people of Te Tiriti o Waitangi, researchers and tangata whenua can collectively make decisions regarding how whakapapa as genomic data is stored and accessed in a mutually beneficial way. For example, as one of few available decapod genomes, the kēkēwai reference genome is likely to be of interest to domestic and international researchers to address both fundamental and applied questions. We are actively engaging with relevant mana whenua regarding the ongoing security and management of these data.

*Concluding Remarks* - We have shown that using a bicultural approach enriches research: In addition to upholding the promises of Te Tiriti o Waitangi, embedding kaupapa Māori principles leads to more contextualised genomic research on taonga species thereby maintaining both the cultural and biological integrity of Aotearoa.

No reira, aukahatia tō waka, kei waiho koe hei tāwai i kā rā o tō oraka.

## Acknowledgements

We are grateful for the support of the Ngāi Tahu Research Centre. We thank Sophie Allen, Greg Brynes, John Hollows for logistical support. We also thank Thomas Buckley, Nick Dunn, Rod Hitchmough, Michael Knapp, Angus McIntosh, Christopher Meijer, Kevin Parker, Mananui Ramsden, Anna Santure, Jeanine Tamati-Elliffe, the SING – Aotearoa 2018 cohort and all members of the Conservation, Evolutionary and Systematics Research Team (ConSERT) for robust dialogue on this topic. We are also grateful for the opportunity provided by the Guest Editors of this Special Issue. This work was funded by the Ministry of Business, Innovation and Employment (New Zealand's Biological Heritage NSC, C09X1501) \*<http://dx.doi.org/10.13039/501100003524>, "Ministry of Business, Innovation and Employment"

## Glossary:

Manaakitanga – respect

Tohungatanga – expertise

Whanaungatanga – relationship, sense of connection

Rangatiratanga – Chieftainship, sovereignty, autonomy, authority

Tikanga – customs, etiquette, protocol

Kaitiakitanga - stewardship

Te Ao Māori – The Māori worldview

Pepeha – tribal saying

Pūrākau – myth, legend, story

Mātauranga Māori – Māori knowledge

Aotearoa – Māori name for New Zealand

Whakapapa – genealogy

Kaupapa – topic, agenda

Waiata – song(s)

Whakataukī – proverbs

Tangata whenua – people of the land

Mahi toi – art

Mauri – life-force

Te Tiriti o Waitangi – The Māori version of The Treaty of Waitangi

Takiwā – territory, area, district

Whānau –family, extended family

## References

Allendorf, F.W.; Hohenlohe, P.A.; Luikart, G. 2010. Genomics and the future of conservation genetics. *Nature reviews genetics* 11: 697.

Ataria, J., Mark-Shadbolt, M., Mead, A.T.P., Prime, K., Doherty, J., Waiwai, J., Ashby, T., Lambert, S. and Garner, G.O. 2018. Whakamanahia Te mātauranga o te Māori: empowering Māori knowledge to support Aotearoa's aquatic biological heritage. *New Zealand Journal of Marine and Freshwater Research*, 52(4), 467-486.

Barrier, R. 2003. *New Zealand Mudfish (Neochanna Spp.) Recovery Plan 2003-13: Northland, Black, Brown, Canterbury, and Chatham Island Mudfish*. Department of Conservation.

Beaton, A.; Hudson, M.; Milne, M.; Port, R.V.; Russell, K.; Smith, B.; Toki, V.; Uerata, L.; Wilcox, P.; Bartholomew, K. 2017. Engaging Māori in biobanking and genomic research: a model for biobanks to guide culturally informed governance, operational, and community engagement activities. *Genetics in Medicine* 19: 345.

Bond, M.O.; Anderson, B.J.; Henare, T.H.A.; Wehi, P.M. 2019. Effects of climatically shifting species distributions on biocultural relationships. *People and Nature* 1: 87-102.

Cadwallader, P.L. 1975. Distribution and ecology of the Canterbury mudfish, *Neochanna burrowsius* (Phillipps)(Salmoniformes: Galaxiidae). *Journal of the Royal Society of New Zealand* 5: 21-30.

Catanach, A.; Crowhurst, R.; Deng, C.; David, C.; Bernatchez, L.; Wellenreuther, M. The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by three-fold in the marine teleost *Chrysophrys auratus*. *Molecular Ecology*. Accepted Author Manuscript

Chen, N. 2019. Conservation: Bye-Bye to the Hihi? *Current Biology* 29: R218-R220.

de Villemereuil, P.; Rutschmann, A.; Lee, K.D.; Ewen, J.G.; Brekke, P.; Santure, A.W. 2019. Little Adaptive Potential in a Threatened Passerine Bird. *Current Biology* 29: 889-894. e883.

Defaveri, J.; Viitaniemi, H.; Leder, E.; Merilä, J. 2013. Characterizing genic and nongenic molecular markers: Comparison of microsatellites and SNPs. *Molecular Ecology Resources* 13: 377-392.

Dunn, N.R.; Allibone, R.M.; Closs, G.; Crow, S.; David, B.O.; Goodman, J.; Griffiths, M.H.; Jack, D.; Ling, N.; Waters, J.M. 2018. *Conservation status of New Zealand freshwater fishes, 2017*. Publishing Team, Department of Conservation.

Frankham, R.; Ballou, J.D.; Briscoe, D.A. 2010. *Introduction to Conservation Genetics*. Cambridge University Press.

Frankham, R.; Ballou, J.D.; Ralls, K.; Eldridge, M.; Dudash, M.R.; Fenster, C.B.; Lacy, R.C.; Sunnucks, P. 2017. *Genetic management of fragmented animal and plant populations*. Oxford University Press.

Funk, W.C.; McKay, J.K.; Hohenlohe, P.A.; Allendorf, F.W. 2012. Harnessing genomics for delineating conservation units. *Trends in ecology & evolution* 27: 489-496.

Galla, S.J.; Buckley, T.R.; Elshire, R.; Hale, M.L.; Knapp, M.; McCallum, J.; Moraga, R.; Santure, A.W.; Wilcox, P.; Steeves, T.E. 2016. Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances. *Molecular ecology* 25: 5267-5281.

Galla, S.J.; Forsdick, N.J.; Brown, L.; Hoepfner, M.; Knapp, M.; Maloney, R.F.; Moraga, R.; Santure, A.W.; Steeves, T.E. 2019. Reference Genomes from Distantly Related Species Can Be Used for Discovery of Single Nucleotide Polymorphisms to Inform Conservation Management. *Genes* 10: 9.

Grainger, N.H., Jon; Drinan, T.C., Kevin; Smith, B.; Death, R.M., Troy; Rolfe, J. 2018. *Conservation status of New Zealand freshwater invertebrates, 2018*. Publishing Team, Department of Conservation.

Harmsworth, G.; Tipa, G. 2006. Māori environmental monitoring in New Zealand: progress, concepts, and future direction. *Report for the ICM website*:

[http://www.landcareresearch.co.nz/research/sustainablesoc/social/indigenous\\_index.asp](http://www.landcareresearch.co.nz/research/sustainablesoc/social/indigenous_index.asp).

Hikuroa, D. 2017. Mātauranga Māori—the ūkaipō of knowledge in New Zealand. *Journal of the Royal Society of New Zealand* 47: 5-10.

Hikuroa, D.; Slade, A.; Gravley, D. 2011. Implementing Māori indigenous knowledge (mātauranga) in a scientific paradigm: Restoring the mauri to Te Kete Poutama. *MAI review* 3: 1-9.

Hoban, S.; Kelley, J.L.; Lotterhos, K.E.; Antolin, M.F.; Bradburd, G.; Lowry, D.B.; Poss, M.L.; Reed, L.K.; Storfer, A.; Whitlock, M.C. 2016. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist* 188: 379-397.

Hudson, M.; Russell, K.; Uerata, L.; Milne, M.; Wilcox, P.; Port, R.V.; Smith, B.; Toki, V.; Beaton, A. 2016. Te Mata Ira—Faces of the Gene: Developing a cultural foundation for biobanking and genomic research involving Māori. *AlterNative: An International Journal of Indigenous Peoples* 12: 341-355.

Hudson, M.L.; Russell, K. 2009. The Treaty of Waitangi and research ethics in Aotearoa. *Journal of Bioethical Inquiry* 6: 61-68.

Kawharu, M. 2000. Kaitiakitanga: a Maori anthropological perspective of the Maori socio-environmental ethic of resource management. *Journal of the Polynesian Society* 109: 349-370.

Kristensen, T.N.; Hoffmann, A.A.; Pertoldi, C.; Stronen, A.V. 2015. What can livestock breeders learn from conservation genetics and vice versa? *Frontiers in genetics* 6: 38.

Kusabs, I.A.; Quinn, J.M. 2009. Use of a traditional Maori harvesting method, the tau kōura, for monitoring kōura (freshwater crayfish, *Paranephrops planifrons*) in Lake Rotoiti, North Island, New Zealand. *New Zealand Journal of Marine and Freshwater Research* 43: 713-722.

Lemopoulos, A.; Prokkola, J.M.; Uusi-Heikkilä, S.; Vasemägi, A.; Huusko, A.; Hyvärinen, P.; Koljonen, M.L.; Koskiniemi, J.; Vainikka, A. 2019. Comparing RADseq and microsatellites for estimating genetic diversity and relatedness—Implications for brown trout conservation. *Ecology and Evolution*.



Lyver, P.O.B.; Ruru, J.; Scott, N.; Tylianakis, J.M.; Arnold, J.; Malinen, S.K.; Bataille, C.Y.; Herse, M.R.; Jones, C.J.; Gormley, A.M. 2018. Building biocultural approaches into Aotearoa–New Zealand’s conservation future. *Journal of the Royal Society of New Zealand*: 1-18.

Lyver, P.O.B.; Tylianakis, J.M. 2017. Indigenous peoples: Conservation paradox. *Science* 357: 142-143.

Mead, H.M. 2003. *Tikanga Māori: Living by Māori Values*. Huia Publishers.

Mercier, O. 2018. Mātauranga and Science. *New Zealand Science Review* 74: 83-90.

Ministry of Research Science and Technology 2007. *Vision Mātauranga: Unlocking the Innovation Potential of Maori Knowledge, Resources and People*. Wellington: Crown Copyright.

Monk, A. 2017. A growing tribal economy. *Te Karaka* 76: 44-46.

O'Brien, L.; Dunn, N. 2007. *Mudfish (Neochanna Galaxiidae) literature review*. Science & Technical Pub., Department of Conservation.

Ouborg, N.J.; Pertoldi, C.; Loeschcke, V.; Bijlsma, R.K.; Hedrick, P.W. 2010. Conservation genetics in transition to conservation genomics. *Trends in genetics* 26: 177-187.

Parkyn, S.; Kusabs, I. 2007. Taonga and mahinga kai species of the Te Arawa lakes: a review of current knowledge–kōura. *NIWA Client Report: HAM2007-022. National Institute of Water and Atmospheric Research, Hamilton, New Zealand*.

Pihama, L. 2012. Kaupapa Māori theory: transforming theory in Aotearoa. *He Pukenga Korero* 9.

Pihama, L.; Cram, F.; Walker, S. 2002. Creating methodological space: A literature review of Kaupapa Maori research. *Canadian Journal of Native Education* 26: 30-43.

Rainforth, H. J. & Harmsworth, G. R. (2019). Kaupapa Māori Freshwater Assessments: A summary of iwi and hapū-based tools, frameworks and methods for assessing freshwater environments.

Perception Planning Ltd. 115 pp.

Reed, A.W. 2004. *Reed book of Māori mythology*. Raupo.

Smith, G.H. 1997. The development of Kaupapa Maori: Theory and praxis. ResearchSpace@ Auckland.

Smith, L.T. 2013. *Decolonizing methodologies: Research and indigenous peoples*. Zed Books Ltd.

Tau, T. 2001. In defence of whakapapa as oral history: a case study. *Te Karaka* 17: 8-9.

Te Rito, J.S. 2007. Whakapapa: A framework for understanding identity. *MAI Review LW* 1: 10.

Te Tiriti o Waitangi 1840. Retrieved from <https://www.waitangitribunal.govt.nz/treaty-of-waitangi/te-reo-maori-version>

Thoms, C. 2016. Distribution, trapping efficiencies and feeding trials for *Paranephrops zealandicus* in central Canterbury [Unpublished MSc thesis]. Christchurch: University of Canterbury.

Walker, R. 1990. *Ka whawhai tonu matou*. Penguin Books.

Walker, S.; Eketone, A.; Gibbs, A. 2006. An exploration of kaupapa Maori research, its principles, processes and applications. *International Journal of Social Research Methodology* 9: 331-344.

Wallis, G.P.; Wallis, L.J. 2014. A Preliminary Transcriptomic Study of Galaxiid Fishes Reveals a Larval Glycoprotein Gene Under Strong Positive Selection. *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, pp 47-68. Springer.

Wehi, P.M.; Cox, M.P.; Roa, T.; Whaanga, H. 2018. Human perceptions of megafaunal extinction events revealed by linguistic analysis of indigenous oral traditions. *Human Ecology* 46: 461-470.

Wehi, P.M.; Lord, J.M. 2017. Importance of including cultural practices in ecological restoration. *Conservation biology* 31: 1109-1118.

Wehi, P.M.; Whaanga, H.; Roa, T. 2009. Missing in translation: Maori language and oral tradition in scientific analyses of traditional ecological knowledge (TEK). *Journal of the Royal Society of New Zealand* 39: 201-204.

Wellenreuther, M.; Le Luyer, J.; Cook, D.; Ritchie, P.A.; Bernatchez, L. 2019. Domestication and temperature modulate gene expression signatures and growth in the Australasian snapper *Chrysophrys auratus*. *G3: Genes, Genomes, Genetics* 9: 105-116.

Whaanga, H.; Wehi, P.; Cox, M.; Roa, T.; Kusabs, I. 2018. Māori oral traditions record and convey indigenous knowledge of marine and freshwater resources. *New Zealand Journal of Marine and Freshwater Research* 52: 487-496.

## Chapter Four: Discussion

### Conservation genomics in Aotearoa New Zealand

The number of genomic research projects on threatened species are expected to increase as genomics becomes more accessible as a tool for conservation. My findings in Chapter Two of this thesis are important for consideration here in Aotearoa New Zealand and beyond. This research provides the first empirical evidence for the use of small sample sizes in population genomic studies of threatened species using whole genome resequencing data. Previously, it was thought that large sample sizes were required for population genetic studies (Yan and Zhang 2004; Ryman et al. 2006; Miyamoto et al. 2008; Pruett and Winter 2008; Hale et al. 2012). Now, there is evidence for the use of smaller sample sizes, which can reduce the cost of population genomic research in threatened species with relatively small populations.

In my thesis I have emphasised the need to consider both sampling design (Chapter Two) and cultural relevance and context (Chapter Three) when establishing a genomic research project on threatened taonga species. I do not expect the approaches outlined in Chapter Three to be directly transferable to all genomics research, however, my co-authors and I hope that our work will provide enough foundation for researchers across Aotearoa to actively develop their own meaningful partnerships with mana whenua to pursue a range of unique, co-developed ideas. Although best-practice is clearly the co-development of research with mana whenua, for those already engaged in genomic research of taonga species, it is never too late to retroactively engage and cultivate new relationships with relevant tribal groups.

## Future directions for the genomics research of kōwaro

This thesis provides a strong foundation for generating genomic data for kōwaro in a culturally responsive way. The obvious future directions for this research are the subsequent generation and analyses of this population genomic data, following the decision-making framework presented in Chapter Three (Figure 3.1). In addition to investigating population structure, kōwaro would benefit from the characterisation of their adaptive variation. Traditionally characterising adaptive variation involved estimating genome-wide diversity through a small number of selectively neutral markers (e.g., microsatellites) and the relationship with fitness was based on the assumption that neutral marker variation is indicative of adaptive variation (Allendorf et al. 2010; Ouborg et al. 2010). However, variation at selectively neutral markers is unlikely to be entirely representative of genome-wide variation and the mechanisms that connect the dynamics of neutral genetic variation and fitness are unknown (DeFaveri et al. 2013). High-throughput sequencing approaches are improving our ability to detect variation at fitness-related loci (Allendorf et al. 2010; Ouborg et al. 2010; Chen 2019; de Villemereuil et al. 2019), which allows us to characterise adaptive potential and use adaptive variation to inform various conservation and management decisions (e.g. Funk et al. 2012).

Detecting adaptive variation can be challenging, and a strong understanding of the species in question can increase the likelihood of detection by informing a robust sampling design that enables the use of genotype – environment association studies (Hoban et al. 2016). Using existing knowledge, *a priori* hypotheses can be formed about the different environmental drivers of selection in kōwaro to ensure that key environmental and ecological gradients are represented within the sampling design to maximise statistical power (Hoban et al. 2016).

In kōwaro, various stressors are having significant effects on demography (e.g. Eldon 1979; Harding et al. 2007; Meijer et al. 2019) and are likely driving selection in different populations. One such example is the variation in drying regimes (Meijer et al. 2019), which could be one of many important abiotic factors linked to selection at fitness-related loci, and therefore may prove to be a

useful ecological measure for characterising adaptive genetic variation. Extensively sampling variation in drying across populations is one strategy that could increase the likelihood of detecting adaptive variation in kōwaro (Hoban et al. 2016). This is one example of the potentially complex ecological and environmental traits that need to be considered for further sampling and analyses.

Novel approaches to analysing genomic data provide a wealth of opportunities for potential research moving forward. For example, going beyond SNP variation at either neutral or selective loci, another potential avenue for kōwaro genomics research is the characterisation of more complex structural genomic variants (Wellenreuther et al. 2019). The small kōwaro genome size ~700 Mb (R.Moraga Unpublished data) provides an opportunity to investigate these kinds of complex variants at a minimal additional cost relative to other vertebrates with larger genomes.

## References

Allendorf, F.W.; Hohenlohe, P.A.; Luikart, G. 2010. Genomics and the future of conservation genetics. *Nature reviews genetics* 11: 697.

Chen, N. 2019. Conservation: Bye-Bye to the Hihi? *Current Biology* 29: R218-R220.

de Villemereuil, P.; Rutschmann, A.; Lee, K.D.; Ewen, J.G.; Brekke, P.; Santure, A.W. 2019. Little Adaptive Potential in a Threatened Passerine Bird. *Current Biology* 29: 889-894. e883.

Defaveri, J.; Viitaniemi, H.; Leder, E.; Merilä, J. 2013. Characterizing genic and nongenic molecular markers: Comparison of microsatellites and SNPs. *Molecular Ecology Resources* 13: 377-392.

Eldon, G. A., 1979. Habitat and interspecific relationships of the Canterbury mudfish, *Neochanna burrowsius* (Salmoniformes: Galaxiidae). *New Zealand journal of marine and freshwater research*, 13(1), 111-119.

Funk, W.C.; McKay, J.K.; Hohenlohe, P.A.; Allendorf, F.W. 2012. Harnessing genomics for delineating conservation units. *Trends in ecology & evolution* 27: 489-496.

Hale, M.L., Burg, T.M. and Steeves, T.E., 2012. Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PLoS one*, 7(9), p.e45170.

Harding, J.S., 2007. Persistence of a significant population of rare Canterbury mudfish (*Neochanna burrowsius*) in a hydrologically isolated catchment. *New Zealand Journal of Marine and Freshwater Research*, 41, pp.309-316.

Hoban, S.; Kelley, J.L.; Lotterhos, K.E.; Antolin, M.F.; Bradburd, G.; Lowry, D.B.; Poss, M.L.; Reed, L.K.; Storfer, A.; Whitlock, M.C. 2016. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist* 188: 379-397.

Meijer, C.G., Warburton, H.J., Harding, J.S. and McIntosh, A.R., 2019. Shifts in population size structure for a drying-tolerant fish in response to extreme drought. *Austral Ecology*.

Miyamoto, N., Fernández-Manjarrés, J.F., Morand-Prieur, M.E., Bertolino, P. and Frascaria-Lacoste, N., 2008. What sampling is needed for reliable estimations of genetic diversity in *Fraxinus excelsior* L.(Oleaceae)?. *Annals of forest science*, 65(4), p.1.

Ouborg, N.J.; Pertoldi, C.; Loeschcke, V.; Bijlsma, R.K.; Hedrick, P.W. 2010. Conservation genetics in transition to conservation genomics. *Trends in genetics* 26: 177-187.

Pruett, C.L. and Winker, K., 2008. The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. *Journal of Avian Biology*, 39(2), pp.252-256.

Wellenreuther, M., Mérot, C., Berdan, E. and Bernatchez, L., 2019. Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. *Molecular ecology*.

Yan, L. and Zhang, D., 2004. Effects of sample size on various genetic diversity measures in population genetic study with microsatellite DNA markers. *Dong wu xue bao.[Acta zoologica Sinica]*, 50(2), pp.279-290.



## Appendix A: Scripts for generating resampled datasets

Bash script for generating Kaki resequencing datasets

```
#!/bin/bash

files="Kaki_5xFinalVariantCalls_BCFTTools_Biallelic_SNPsOnly_MAF0.05_AVGDP_Q20_LD0.8_Missing0.1_thin15
0.vcf.recode.vcf

Kaki_9xFinalVariantCalls_BCFTTools_Biallelic_SNPsOnly_MAF0.05_AVGDP_Q20_LD0.8_Missing0.1_thin150.vcf.r
ecode.vcf

Kaki_FinalVariantCalls_BCFTTools_Biallelic_SNPsOnly_MAF0.05_AVGDP_Q20_LD0.8_Missing0.1_thin150.vcf.rec
ode.vcf"

for (( n=1; n<=1000; n++))

do cd /media/levi/1TB/Kaki_reseq/

    for vcf in $files

    do

        for i in 2 4 6 8 10 12 14 16 18 20 22 24

        do vcftools --vcf $vcf --max-indv $i --out

/media/levi/1TB/Kaki_reseq/Sub5_9x/$i\Indiv_${x%%FinalVariantCalls_BCFTTools_Biallelic_SNPsOnly_MAF0.05
_AVGDP_Q20_LD0.8_Missing0.1_thin150.vcf.recode.vcf} --recode

        done

    done

done
```

```

    for vcf in

$Kaki_FinalVariantCalls_BCFTTools_Biallelic_SNPsOnly_MAF0.05_AVGDP_Q20_LD0.8_Missing0.1_thin150.vcf.r
ecode.vcf

    do

        for i in 26 28 30 32 34

            do vcftools --vcf $vcf --max-indv $i --out

/media/levi/1TB/Kaki_reseq/Sub5_9x/$i\Indiv_${x%%FinalVariantCalls_BCFTTools_Biallelic_SNPsOnly_MAF0.05
_AVGDP_Q20_LD0.8_Missing0.1_thin150.vcf.recode.vcf} --recode

            done

        done

        echo "subsamped vcfs for individuals $n"

        cd /media/levi/1TB/Kaki_reseq/Sub5_9x

        for vcf in *.vcf

            do perl /media/levi/1TB/calculate_stats_VCF.pl $vcf >>

/media/levi/1TB/Kaki_reseq/Subsample_CSVs/${vcf%.recode.vcf}.csv

            done

        echo "generated statistics for individuals $n"

done

```

```
cd /media/levi/1TB/Kaki_reseq/Subsample_CSVs
```

```
for csv in *
```

```
do cat /media/levi/1TB/Header.csv $csv > /media/levi/1TB/Kaki_reseq/csvs_with_headers/$csv
```

```
done
```

[Bash script for generating Kakī GBS datasets](#)

```
#!/bin/bash
```

```
subs=/media/levi/1TB/Subsampled_VCFs
```

```
files="2016_2018_10%.vcf
```

```
2016_2018_20%.vcf
```

```
2016_2018_30%.vcf
```

```
2016_2018_40%.vcf
```

```
2016_2018_50%.vcf
```

```
2016_2018_60%.vcf
```

```
2016_2018_70%.vcf
```

```
2016_2018_80%.vcf
```

```
2016_2018_90%.vcf
```

```
2016_2018_100%.vcf"
```

```
for (( n=1; n<=1000; n++))
```

```
do cd $subs
```

```
    for vcf in $files
```

```
    do name=${vcf##2016_2018_}
```

```
        for x in 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 50 60 70 80 90 100
```

```
        do perl /media/levi/1TB/subsample_VCF_columns.pl $vcf >
```

```
subsample_individuals/$x%Indiv_${name}%.vcf)Depth.vcf $x
```

```
        done
```

```
    done
```

```
echo "subsampled vcfs for individuals $n"
```

```
cd $subs/subsample_individuals
```

```
    for sample in *.vcf
```

```
    do perl /media/levi/1TB/calculate_stats_VCF.pl $sample >>
```

```
/media/levi/1TB/Subsampled_VCFs/depth_csvs/${sample%.vcf}.csv
```

```
    done
```

```
echo "generated statistics for individuals $n"
```

```
done
```

```
cd $subs/depth_csvs
```

```
for csv in *
```

```
do cat /media/levi/1TB/Header.csv $csv > /media/levi/1TB/Subsampled_VCFs/csvs_with_headers/$csv
```

```
done
```

Bash script for generating Bullers' Albatross resampling datasets

```
#!/bin/bash
```

```
subs=/media/levi/1TB/Bullers/Subsampled_VCFs
```

```
files="bullers_all.vcf
```

```
bullers_N.vcf
```

```
bullers_S.vcf"
```

```
for (( n=1; n<=300; n++))
```

```
do cd $subs
```

```
    for vcf in $files
```

```
    do
```

```
        for x in 2 4 6 8 10 12 14 16 18 20 22 24 26 28
```

```
        do vcftools --vcf $vcf --max-indv $x --out $subs/subsample_individuals/$x\Indiv_${vcf%%.vcf}
```

```
--recode
```

```
        done
```

done

echo "subsamped vcfs for individuals \$n"

cd \$subs/subsample\_individuals

for sample in \*.vcf

do perl /media/levi/1TB/calculate\_stats\_VCF.pl \$sample >>

\$subs/depth\_csvs/\${sample%%.recode.vcf}.csv

done

echo "generated statistics for individuals \$n"

done

cd \$subs/depth\_csvs

for csv in \*

do cat /media/levi/1TB/Header.csv \$csv > /media/levi/1TB/Subsampled\_VCFs/csvs\_with\_headers/\$csv

done

## Appendix B: R code for resampled datasets

### R script for Resequencing data

```
library(dplyr)

setwd("/media/levi/1TB/Kaki_reseq/csvs_with_headers")

#importing subsampling csvs

#adding column for number of individuals

Indiv34_Reseq<-read.csv("34Indiv.csv") %>% mutate(Individuals = rep(34,n()))

Indiv32_Reseq<-read.csv("32Indiv.csv") %>% mutate(Individuals = rep(32,n()))

Indiv30_Reseq<-read.csv("30Indiv.csv") %>% mutate(Individuals = rep(30,n()))

Indiv28_Reseq<-read.csv("28Indiv.csv") %>% mutate(Individuals = rep(28,n()))

Indiv26_Reseq<-read.csv("26Indiv.csv") %>% mutate(Individuals = rep(26,n()))

Indiv24_Reseq<-read.csv("24Indiv.csv") %>% mutate(Individuals = rep(24,n()))

Indiv22_Reseq<-read.csv("22Indiv.csv") %>% mutate(Individuals = rep(22,n()))

Indiv20_Reseq<-read.csv("20Indiv.csv") %>% mutate(Individuals = rep(20,n()))

Indiv18_Reseq<-read.csv("18Indiv.csv") %>% mutate(Individuals = rep(18,n()))

Indiv16_Reseq<-read.csv("16Indiv.csv") %>% mutate(Individuals = rep(16,n()))

Indiv14_Reseq<-read.csv("14Indiv.csv") %>% mutate(Individuals = rep(14,n()))

Indiv12_Reseq<-read.csv("12Indiv.csv") %>% mutate(Individuals = rep(12,n()))

Indiv10_Reseq<-read.csv("10Indiv.csv") %>% mutate(Individuals = rep(10,n()))

Indiv08_Reseq<-read.csv("08Indiv.csv") %>% mutate(Individuals = rep(8,n()))
```

```
Indiv06_Reseq<-read.csv("06Indiv.csv") %>% mutate(Individuals = rep(6,n()))
```

```
Indiv04_Reseq<-read.csv("04Indiv.csv") %>% mutate(Individuals = rep(4,n()))
```

```
Indiv02_Reseq<-read.csv("02Indiv.csv") %>% mutate(Individuals = rep(2,n()))
```

```
#combining to create file with all subsamples together
```

```
combinedreseq<-
```

```
bind_rows(Indiv02_Reseq,Indiv04_Reseq,Indiv06_Reseq,Indiv08_Reseq,Indiv10_Reseq,Indiv12_Reseq,Indiv14  
_Reseq,Indiv16_Reseq,Indiv18_Reseq,Indiv20_Reseq,Indiv22_Reseq,Indiv24_Reseq,Indiv26_Reseq,Indiv28_Re  
seq,Indiv30_Reseq,Indiv32_Reseq,Indiv34_Reseq)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP<-combinedreseq %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value,  
Individuals)
```

```
combinedHetSNP<-combinedreseq %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs =  
value, Individuals)
```

```
combinedHetENT<-combinedreseq %>% filter(stat == "Total Heterozygous entries min. 3") %>%  
select(HetENTs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```

```
SubReseq<-na.omit(bind_cols(combinedHetSNP, combinedTotSNP, combinedHetENT)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs, HetENTs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs) %>%
```

```
group_by(Individuals) %>%
```



```
filter(row_number())>=(n()-999))
```

```
#Creating an object containing mean value per no. Individuals
```

```
AVG<-SubReseq %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
    HetSNPs = mean(HetSNPs),
```

```
    TotSNPs = mean(TotSNPs),
```

```
    PropSNPs = mean(PropSNPs),
```

```
    HetENTs = mean(HetENTs))
```

```
#importing subsampling csvs
```

```
#adding column for number of individuals
```

```
Indiv24_Reseq9x<-read.csv("24Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(24,n()))
```

```
Indiv22_Reseq9x<-read.csv("22Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(22,n()))
```

```
Indiv20_Reseq9x<-read.csv("20Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(20,n()))
```

```
Indiv18_Reseq9x<-read.csv("18Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(18,n()))
```

```
Indiv16_Reseq9x<-read.csv("16Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(16,n()))
```

```
Indiv14_Reseq9x<-read.csv("14Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(14,n()))
```

```
Indiv12_Reseq9x<-read.csv("12Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(12,n()))
```

```
Indiv10_Reseq9x<-read.csv("10Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(10,n()))
```

```
Indiv08_Reseq9x<-read.csv("8Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(8,n()))
```

```
Indiv06_Reseq9x<-read.csv("6Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(6,n()))
```

```
Indiv04_Reseq9x<-read.csv("4Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(4,n()))
```

```
Indiv02_Reseq9x<-read.csv("2Indiv_Kaki_9x.csv") %>% mutate(Individuals = rep(2,n()))
```

```
#combining to create file with all subsamples together
```

```
combined9xreseq<-
```

```
bind_rows(Indiv02_Reseq9x,Indiv04_Reseq9x,Indiv06_Reseq9x,Indiv08_Reseq9x,Indiv10_Reseq9x,Indiv12_Reseq9x,Indiv14_Reseq9x,Indiv16_Reseq9x,Indiv18_Reseq9x,Indiv20_Reseq9x,Indiv22_Reseq9x,Indiv24_Reseq9x)
```

```
#creating subsets with each of the relevant statistics
```

```
combined9xTotSNP<-combined9xreseq %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
combined9xHetSNP<-combined9xreseq %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
combined9xHetENT<-combined9xreseq %>% filter(stat == "Total Heterozygous entries min. 3") %>% select(HetENTs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of heterozygous SNPs
```

```
SubReseq9x<-na.omit(bind_cols(combined9xHetSNP, combined9xTotSNP, combined9xHetENT)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs,HetENTs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number())>=(n()-999))
```

```
#Creating an object containing mean value per no. Individuals
```

```
AVG9x<-SubReseq9x %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
  HetSNPs = mean(HetSNPs),
```

```
  TotSNPs = mean(TotSNPs),
```

```
  PropSNPs = mean(PropSNPs),
```

```
  HetENTs = mean(HetENTs))
```

```
#importing subsampling csvs
```

```
#adding column for number of individuals
```

```
Indiv24_Reseq5x<-read.csv("24Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(24,n()))
```

```
Indiv22_Reseq5x<-read.csv("22Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(22,n()))
```

```
Indiv20_Reseq5x<-read.csv("20Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(20,n()))
```

```
Indiv18_Reseq5x<-read.csv("18Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(18,n()))
```

```
Indiv16_Reseq5x<-read.csv("16Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(16,n()))
```

```
Indiv14_Reseq5x<-read.csv("14Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(14,n()))
```

```
Indiv12_Reseq5x<-read.csv("12Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(12,n()))
```

```
Indiv10_Reseq5x<-read.csv("10Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(10,n()))
```

```
Indiv08_Reseq5x<-read.csv("8Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(8,n()))
```

```
Indiv06_Reseq5x<-read.csv("6Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(6,n()))
```

```
Indiv04_Reseq5x<-read.csv("4Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(4,n()))
```

```
Indiv02_Reseq5x<-read.csv("2Indiv_Kaki_5x.csv") %>% mutate(Individuals = rep(2,n()))
```

```
#combining to create file with all subsamples together
```

```
combined5xreseq<-
```

```
bind_rows(Indiv02_Reseq5x,Indiv04_Reseq5x,Indiv06_Reseq5x,Indiv08_Reseq5x,Indiv10_Reseq5x,Indiv12_Reseq5x,Indiv14_Reseq5x,Indiv16_Reseq5x,Indiv18_Reseq5x,Indiv20_Reseq5x,Indiv22_Reseq5x,Indiv24_Reseq5x)
```

```
#creating subsets with each of the relevant statistics
```

```
combined5xTotSNP<-combined5xreseq %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
combined5xHetSNP<-combined5xreseq %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
combined5xHetENT<-combined5xreseq %>% filter(stat == "Total Heterozygous entries min. 3") %>% select(HetENTs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```

```
SubReseq5x<-na.omit(bind_cols(combined5xHetSNP, combined5xTotSNP, combined5xHetENT)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs, HetENTs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number()>=(n()-999))
```

```
#Creating an object containing mean value per no. Individuals
```

```
AVG5x<-SubReseq5x %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
  HetSNPs = mean(HetSNPs),
```

```
  TotSNPs = mean(TotSNPs),
```

```
  PropSNPs = mean(PropSNPs),
```

```
  HetENTs = mean(HetENTs))
```

[R script for Kakī GBS data](#)

```
library(dplyr)
```

```
setwd("/media/levi/1TB/Subsampled_VCFs/csvs_with_headers")
```

```
#
```

```
#Importing All stat observations for different %Individuals at 100% Depth of Coverage
```

```
Indiv100_100Depth<-read.csv("100%Indiv_100Depth.csv") %>% mutate(Individuals = rep(100,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv90_100Depth<-read.csv("90%Indiv_100Depth.csv") %>% mutate(Individuals = rep(90,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv80_100Depth<-read.csv("80%Indiv_100Depth.csv") %>% mutate(Individuals = rep(80,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv70_100Depth<-read.csv("70%Indiv_100Depth.csv") %>% mutate(Individuals = rep(70,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv60_100Depth<-read.csv("60%Indiv_100Depth.csv") %>% mutate(Individuals = rep(60,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv50_100Depth<-read.csv("50%Indiv_100Depth.csv") %>% mutate(Individuals = rep(50,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv40_100Depth<-read.csv("40%Indiv_100Depth.csv") %>% mutate(Individuals = rep(40,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv38_100Depth<-read.csv("38%Indiv_100Depth.csv") %>% mutate(Individuals = rep(38,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv36_100Depth<-read.csv("36%Indiv_100Depth.csv") %>% mutate(Individuals = rep(36,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv34_100Depth<-read.csv("34%Indiv_100Depth.csv") %>% mutate(Individuals = rep(34,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv32_100Depth<-read.csv("32%Indiv_100Depth.csv") %>% mutate(Individuals = rep(32,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv30_100Depth<-read.csv("30%Indiv_100Depth.csv") %>% mutate(Individuals = rep(30,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv28_100Depth<-read.csv("28%Indiv_100Depth.csv") %>% mutate(Individuals = rep(28,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv26_100Depth<-read.csv("26%Indiv_100Depth.csv") %>% mutate(Individuals = rep(26,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv24_100Depth<-read.csv("24%Indiv_100Depth.csv") %>% mutate(Individuals = rep(24,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv22_100Depth<-read.csv("22%Indiv_100Depth.csv") %>% mutate(Individuals = rep(22,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv20_100Depth<-read.csv("20%Indiv_100Depth.csv") %>% mutate(Individuals = rep(20,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv18_100Depth<-read.csv("18%Indiv_100Depth.csv") %>% mutate(Individuals = rep(18,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv16_100Depth<-read.csv("16%Indiv_100Depth.csv") %>% mutate(Individuals = rep(16,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv14_100Depth<-read.csv("14%Indiv_100Depth.csv") %>% mutate(Individuals = rep(14,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv12_100Depth<-read.csv("12%Indiv_100Depth.csv") %>% mutate(Individuals = rep(12,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv10_100Depth<-read.csv("10%Indiv_100Depth.csv") %>% mutate(Individuals = rep(10,n())) %>% slice(-  
5731:-6180) %>% slice(-6433:-7608)
```

```
Indiv08_100Depth<-read.csv("8%Indiv_100Depth.csv") %>% mutate(Individuals = rep(8,n())) %>% slice(-5731:-  
6180) %>% slice(-6433:-7608)
```

```
Indiv06_100Depth<-read.csv("6%Indiv_100Depth.csv") %>% mutate(Individuals = rep(6,n())) %>% slice(-5731:-  
6180) %>% slice(-6433:-7608)
```

```
Indiv04_100Depth<-read.csv("4%Indiv_100Depth.csv") %>% mutate(Individuals = rep(4,n())) %>% slice(-5731:-  
6180) %>% slice(-6433:-7608)
```

```
Indiv02_100Depth<-read.csv("2%Indiv_100Depth.csv") %>% mutate(Individuals = rep(2,n())) %>% slice(-5731:-  
6180) %>% slice(-6433:-7608)
```

```
#combine all observations
```

```
combined100<-
```

```
bind_rows(Indiv02_100Depth,Indiv04_100Depth,Indiv06_100Depth,Indiv08_100Depth,Indiv10_100Depth,Indiv12_100Depth,Indiv14_100Depth,Indiv16_100Depth,Indiv18_100Depth,Indiv20_100Depth,Indiv22_100Depth,Indiv24_100Depth,Indiv26_100Depth,Indiv28_100Depth,Indiv30_100Depth,Indiv32_100Depth,Indiv34_100Depth,Indiv36_100Depth,Indiv38_100Depth,Indiv40_100Depth,Indiv50_100Depth,Indiv60_100Depth,Indiv70_100Depth,Indiv80_100Depth,Indiv90_100Depth,Indiv100_100Depth)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP100<-combined100 %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
combinedHetSNP100<-combined100 %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```

```
Sub100<-na.omit(bind_cols(combinedHetSNP100, combinedTotSNP100)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs, Individuals = (Individuals/100)*88) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number())>=(n()-999))
```

```
#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals (because some observations are lost as NA values during data generation)
```

```
AVG100<-Sub100 %>%
```



```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
    HetSNPs = mean(HetSNPs),
```

```
    TotSNPs = mean(TotSNPs),
```

```
    PropSNPs = mean(PropSNPs))
```

```
#
```

```
#
```

```
#Importing All stat observations for different %Individuals at 90% Depth of Coverage
```

```
Indiv100_90Depth<-read.csv("100%Indiv_90Depth.csv") %>% mutate(Individuals = rep(100,n())) %>% slice(-  
2257:-2646)
```

```
Indiv90_90Depth<-read.csv("90%Indiv_90Depth.csv") %>% mutate(Individuals = rep(90,n())) %>% slice(-2257:-  
2646)
```

```
Indiv80_90Depth<-read.csv("80%Indiv_90Depth.csv") %>% mutate(Individuals = rep(80,n())) %>% slice(-2257:-  
2646)
```

```
Indiv70_90Depth<-read.csv("70%Indiv_90Depth.csv") %>% mutate(Individuals = rep(70,n())) %>% slice(-2257:-  
2646)
```

```
Indiv60_90Depth<-read.csv("60%Indiv_90Depth.csv") %>% mutate(Individuals = rep(60,n())) %>% slice(-2257:-  
2646)
```

```
Indiv50_90Depth<-read.csv("50%Indiv_90Depth.csv") %>% mutate(Individuals = rep(50,n())) %>% slice(-2257:-  
2646)
```

```
Indiv40_90Depth<-read.csv("40%Indiv_90Depth.csv") %>% mutate(Individuals = rep(40,n())) %>% slice(-2257:-  
2646)
```

```
Indiv38_90Depth<-read.csv("38%Indiv_90Depth.csv") %>% mutate(Individuals = rep(38,n())) %>% slice(-2257:-  
2646)
```

```
Indiv36_90Depth<-read.csv("36%Indiv_90Depth.csv") %>% mutate(Individuals = rep(36,n())) %>% slice(-2257:-  
2646)
```

```
Indiv34_90Depth<-read.csv("34%Indiv_90Depth.csv") %>% mutate(Individuals = rep(34,n())) %>% slice(-2257:-  
2646)
```

```
Indiv32_90Depth<-read.csv("32%Indiv_90Depth.csv") %>% mutate(Individuals = rep(32,n())) %>% slice(-2257:-  
2646)
```

```
Indiv30_90Depth<-read.csv("30%Indiv_90Depth.csv") %>% mutate(Individuals = rep(30,n())) %>% slice(-2257:-  
2646)
```

```
Indiv28_90Depth<-read.csv("28%Indiv_90Depth.csv") %>% mutate(Individuals = rep(28,n())) %>% slice(-2257:-  
2646)
```

```
Indiv26_90Depth<-read.csv("26%Indiv_90Depth.csv") %>% mutate(Individuals = rep(26,n())) %>% slice(-2257:-  
2646)
```

```
Indiv24_90Depth<-read.csv("24%Indiv_90Depth.csv") %>% mutate(Individuals = rep(24,n())) %>% slice(-2257:-  
2646)
```

```
Indiv22_90Depth<-read.csv("22%Indiv_90Depth.csv") %>% mutate(Individuals = rep(22,n())) %>% slice(-2257:-  
2646)
```

```
Indiv20_90Depth<-read.csv("20%Indiv_90Depth.csv") %>% mutate(Individuals = rep(20,n())) %>% slice(-2257:-  
2646)
```

```
Indiv18_90Depth<-read.csv("18%Indiv_90Depth.csv") %>% mutate(Individuals = rep(18,n())) %>% slice(-2257:-  
2646)
```

```
Indiv16_90Depth<-read.csv("16%Indiv_90Depth.csv") %>% mutate(Individuals = rep(16,n())) %>% slice(-2257:-  
2646)
```

```
Indiv14_90Depth<-read.csv("14%Indiv_90Depth.csv") %>% mutate(Individuals = rep(14,n())) %>% slice(-2257:-  
2646)
```

```
Indiv12_90Depth<-read.csv("12%Indiv_90Depth.csv") %>% mutate(Individuals = rep(12,n())) %>% slice(-2257:-  
2646)
```

```
Indiv10_90Depth<-read.csv("10%Indiv_90Depth.csv") %>% mutate(Individuals = rep(10,n())) %>% slice(-2257:-  
2646)
```

```
Indiv08_90Depth<-read.csv("8%Indiv_90Depth.csv") %>% mutate(Individuals = rep(8,n())) %>% slice(-2257:-  
2646)
```

```
Indiv06_90Depth<-read.csv("6%Indiv_90Depth.csv") %>% mutate(Individuals = rep(6,n())) %>% slice(-2257:-  
2646)
```

```
Indiv04_90Depth<-read.csv("4%Indiv_90Depth.csv") %>% mutate(Individuals = rep(4,n())) %>% slice(-2257:-  
2646)
```

```
Indiv02_90Depth<-read.csv("2%Indiv_90Depth.csv") %>% mutate(Individuals = rep(2,n())) %>% slice(-2257:-  
2646)
```

```
#combine all observations
```

```
combined90<-
```

```
bind_rows(Indiv02_90Depth,Indiv04_90Depth,Indiv06_90Depth,Indiv08_90Depth,Indiv10_90Depth,Indiv12_9  
0Depth,Indiv14_90Depth,Indiv16_90Depth,Indiv18_90Depth,Indiv20_90Depth,Indiv22_90Depth,Indiv24_90D  
epth,Indiv26_90Depth,Indiv28_90Depth,Indiv30_90Depth,Indiv32_90Depth,Indiv34_90Depth,Indiv36_90Dept  
h,Indiv38_90Depth,Indiv40_90Depth,Indiv50_90Depth,Indiv60_90Depth,Indiv70_90Depth,Indiv80_90Depth,In  
div90_90Depth,Indiv100_90Depth)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP90<-combined90 %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value,  
Individuals)
```

```
combinedHetSNP90<-combined90 %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs =  
value, Individuals)
```

#combining these and then creating an additional column containing the proportion of SNPs

```
Sub90<-na.omit(bind_cols(combinedHetSNP90, combinedTotSNP90)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs, Individuals = (Individuals/100)*88) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number())>=(n()-999))
```

#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals

```
AVG90<-Sub90 %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
  HetSNPs = mean(HetSNPs),
```

```
  TotSNPs = mean(TotSNPs),
```

```
  PropSNPs = mean(PropSNPs))
```

```
#
```

#Importing All stat observations for different %Individuals at 80% Depth of Coverage

```
Indiv100_80Depth<-read.csv("100%Indiv_80Depth.csv") %>% mutate(Individuals = rep(100,n())) %>% slice(-  
2257:-2646)
```

```
Indiv90_80Depth<-read.csv("90%Indiv_80Depth.csv") %>% mutate(Individuals = rep(90,n())) %>% slice(-2257:-  
2646)
```

```
Indiv80_80Depth<-read.csv("80%Indiv_80Depth.csv") %>% mutate(Individuals = rep(80,n())) %>% slice(-2257:-  
2646)
```

```
Indiv70_80Depth<-read.csv("70%Indiv_80Depth.csv") %>% mutate(Individuals = rep(70,n())) %>% slice(-2257:-  
2646)
```

```
Indiv60_80Depth<-read.csv("60%Indiv_80Depth.csv") %>% mutate(Individuals = rep(60,n())) %>% slice(-2257:-  
2646)
```

```
Indiv50_80Depth<-read.csv("50%Indiv_80Depth.csv") %>% mutate(Individuals = rep(50,n())) %>% slice(-2257:-  
2646)
```

```
Indiv40_80Depth<-read.csv("40%Indiv_80Depth.csv") %>% mutate(Individuals = rep(40,n())) %>% slice(-2257:-  
2646)
```

```
Indiv38_80Depth<-read.csv("38%Indiv_80Depth.csv") %>% mutate(Individuals = rep(38,n())) %>% slice(-2257:-  
2646)
```

```
Indiv36_80Depth<-read.csv("36%Indiv_80Depth.csv") %>% mutate(Individuals = rep(36,n())) %>% slice(-2257:-  
2646)
```

```
Indiv34_80Depth<-read.csv("34%Indiv_80Depth.csv") %>% mutate(Individuals = rep(34,n())) %>% slice(-2257:-  
2646)
```

```
Indiv32_80Depth<-read.csv("32%Indiv_80Depth.csv") %>% mutate(Individuals = rep(32,n())) %>% slice(-2257:-  
2646)
```

```
Indiv30_80Depth<-read.csv("30%Indiv_80Depth.csv") %>% mutate(Individuals = rep(30,n())) %>% slice(-2257:-  
2646)
```

```
Indiv28_80Depth<-read.csv("28%Indiv_80Depth.csv") %>% mutate(Individuals = rep(28,n())) %>% slice(-2257:-  
2646)
```

```
Indiv26_80Depth<-read.csv("26%Indiv_80Depth.csv") %>% mutate(Individuals = rep(26,n())) %>% slice(-2257:-  
2646)
```

```
Indiv24_80Depth<-read.csv("24%Indiv_80Depth.csv") %>% mutate(Individuals = rep(24,n())) %>% slice(-2257:-  
2646)
```

```
Indiv22_80Depth<-read.csv("22%Indiv_80Depth.csv") %>% mutate(Individuals = rep(22,n())) %>% slice(-2257:-  
2646)
```

```
Indiv20_80Depth<-read.csv("20%Indiv_80Depth.csv") %>% mutate(Individuals = rep(20,n())) %>% slice(-2257:-  
2646)
```

```
Indiv18_80Depth<-read.csv("18%Indiv_80Depth.csv") %>% mutate(Individuals = rep(18,n())) %>% slice(-2257:-  
2646)
```

```
Indiv16_80Depth<-read.csv("16%Indiv_80Depth.csv") %>% mutate(Individuals = rep(16,n())) %>% slice(-2257:-  
2646)
```

```
Indiv14_80Depth<-read.csv("14%Indiv_80Depth.csv") %>% mutate(Individuals = rep(14,n())) %>% slice(-2257:-  
2646)
```

```
Indiv12_80Depth<-read.csv("12%Indiv_80Depth.csv") %>% mutate(Individuals = rep(12,n())) %>% slice(-2257:-  
2646)
```

```
Indiv10_80Depth<-read.csv("10%Indiv_80Depth.csv") %>% mutate(Individuals = rep(10,n())) %>% slice(-2257:-  
2646)
```

```
Indiv08_80Depth<-read.csv("8%Indiv_80Depth.csv") %>% mutate(Individuals = rep(8,n())) %>% slice(-2257:-  
2646)
```

```
Indiv06_80Depth<-read.csv("6%Indiv_80Depth.csv") %>% mutate(Individuals = rep(6,n())) %>% slice(-2257:-  
2646)
```

```
Indiv04_80Depth<-read.csv("4%Indiv_80Depth.csv") %>% mutate(Individuals = rep(4,n())) %>% slice(-2257:-2646)
```

```
Indiv02_80Depth<-read.csv("2%Indiv_80Depth.csv") %>% mutate(Individuals = rep(2,n())) %>% slice(-2257:-2646)
```

```
#combine all observations
```

```
combined80<-
```

```
bind_rows(Indiv02_80Depth,Indiv04_80Depth,Indiv06_80Depth,Indiv08_80Depth,Indiv10_80Depth,Indiv12_80Depth,Indiv14_80Depth,Indiv16_80Depth,Indiv18_80Depth,Indiv20_80Depth,Indiv22_80Depth,Indiv24_80Depth,Indiv26_80Depth,Indiv28_80Depth,Indiv30_80Depth,Indiv32_80Depth,Indiv34_80Depth,Indiv36_80Depth,Indiv38_80Depth,Indiv40_80Depth,Indiv50_80Depth,Indiv60_80Depth,Indiv70_80Depth,Indiv80_80Depth,Indiv90_80Depth,Indiv100_80Depth)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP80<-combined80 %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
combinedHetSNP80<-combined80 %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```

```
Sub80<-na.omit(bind_cols(combinedHetSNP80, combinedTotSNP80)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs, Individuals = (Individuals/100)*88) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number())>=(n()-999))
```

```
#Creating an object containing mean value per no. Individuals and a count of observations for each number of  
individuals
```

```
AVG80<-Sub80 %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
    HetSNPs = mean(HetSNPs),
```

```
    TotSNPs = mean(TotSNPs),
```

```
    PropSNPs = mean(PropSNPs))
```

```
#
```

```
#
```

```
#Importing All stat observations for different %Individuals at 70% Depth of Coverage
```

```
Indiv100_70Depth<-read.csv("100%Indiv_70Depth.csv") %>% mutate(Individuals = rep(100,n())) %>% slice(-  
2257:-2646)
```

```
Indiv90_70Depth<-read.csv("90%Indiv_70Depth.csv") %>% mutate(Individuals = rep(90,n())) %>% slice(-2257:-  
2646)
```

```
Indiv80_70Depth<-read.csv("80%Indiv_70Depth.csv") %>% mutate(Individuals = rep(80,n())) %>% slice(-2257:-  
2646)
```

```
Indiv70_70Depth<-read.csv("70%Indiv_70Depth.csv") %>% mutate(Individuals = rep(70,n())) %>% slice(-2257:-  
2646)
```



```
Indiv60_70Depth<-read.csv("60%Indiv_70Depth.csv") %>% mutate(Individuals = rep(60,n())) %>% slice(-2257:-  
2646)
```

```
Indiv50_70Depth<-read.csv("50%Indiv_70Depth.csv") %>% mutate(Individuals = rep(50,n())) %>% slice(-2257:-  
2646)
```

```
Indiv40_70Depth<-read.csv("40%Indiv_70Depth.csv") %>% mutate(Individuals = rep(40,n())) %>% slice(-2257:-  
2646)
```

```
Indiv38_70Depth<-read.csv("38%Indiv_70Depth.csv") %>% mutate(Individuals = rep(38,n())) %>% slice(-2257:-  
2646)
```

```
Indiv36_70Depth<-read.csv("36%Indiv_70Depth.csv") %>% mutate(Individuals = rep(36,n())) %>% slice(-2257:-  
2646)
```

```
Indiv34_70Depth<-read.csv("34%Indiv_70Depth.csv") %>% mutate(Individuals = rep(34,n())) %>% slice(-2257:-  
2646)
```

```
Indiv32_70Depth<-read.csv("32%Indiv_70Depth.csv") %>% mutate(Individuals = rep(32,n())) %>% slice(-2257:-  
2646)
```

```
Indiv30_70Depth<-read.csv("30%Indiv_70Depth.csv") %>% mutate(Individuals = rep(30,n())) %>% slice(-2257:-  
2646)
```

```
Indiv28_70Depth<-read.csv("28%Indiv_70Depth.csv") %>% mutate(Individuals = rep(28,n())) %>% slice(-2257:-  
2646)
```

```
Indiv26_70Depth<-read.csv("26%Indiv_70Depth.csv") %>% mutate(Individuals = rep(26,n())) %>% slice(-2257:-  
2646)
```

```
Indiv24_70Depth<-read.csv("24%Indiv_70Depth.csv") %>% mutate(Individuals = rep(24,n())) %>% slice(-2257:-  
2646)
```

```
Indiv22_70Depth<-read.csv("22%Indiv_70Depth.csv") %>% mutate(Individuals = rep(22,n())) %>% slice(-2257:-  
2646)
```

```
Indiv20_70Depth<-read.csv("20%Indiv_70Depth.csv") %>% mutate(Individuals = rep(20,n())) %>% slice(-2257:-  
2646)
```

```
Indiv18_70Depth<-read.csv("18%Indiv_70Depth.csv") %>% mutate(Individuals = rep(18,n())) %>% slice(-2257:-  
2646)
```

```
Indiv16_70Depth<-read.csv("16%Indiv_70Depth.csv") %>% mutate(Individuals = rep(16,n())) %>% slice(-2257:-  
2646)
```

```
Indiv14_70Depth<-read.csv("14%Indiv_70Depth.csv") %>% mutate(Individuals = rep(14,n())) %>% slice(-2257:-  
2646)
```

```
Indiv12_70Depth<-read.csv("12%Indiv_70Depth.csv") %>% mutate(Individuals = rep(12,n())) %>% slice(-2257:-  
2646)
```

```
Indiv10_70Depth<-read.csv("10%Indiv_70Depth.csv") %>% mutate(Individuals = rep(10,n())) %>% slice(-2257:-  
2646)
```

```
Indiv08_70Depth<-read.csv("8%Indiv_70Depth.csv") %>% mutate(Individuals = rep(8,n())) %>% slice(-2257:-  
2646)
```

```
Indiv06_70Depth<-read.csv("6%Indiv_70Depth.csv") %>% mutate(Individuals = rep(6,n())) %>% slice(-2257:-  
2646)
```

```
Indiv04_70Depth<-read.csv("4%Indiv_70Depth.csv") %>% mutate(Individuals = rep(4,n())) %>% slice(-2257:-  
2646)
```

```
Indiv02_70Depth<-read.csv("2%Indiv_70Depth.csv") %>% mutate(Individuals = rep(2,n())) %>% slice(-2257:-  
2646)
```

```
#combine all observations
```

```
combined70<-
```

```
bind_rows(Indiv02_70Depth,Indiv04_70Depth,Indiv06_70Depth,Indiv08_70Depth,Indiv10_70Depth,Indiv12_7  
0Depth,Indiv14_70Depth,Indiv16_70Depth,Indiv18_70Depth,Indiv20_70Depth,Indiv22_70Depth,Indiv24_70D
```

```
epth,Indiv26_70Depth,Indiv28_70Depth,Indiv30_70Depth,Indiv32_70Depth,Indiv34_70Depth,Indiv36_70Depth,Indiv38_70Depth,Indiv40_70Depth,Indiv50_70Depth,Indiv60_70Depth,Indiv70_70Depth,Indiv80_70Depth,Indiv90_70Depth,Indiv100_70Depth)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP70<-combined70 %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
combinedHetSNP70<-combined70 %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs, removing all NA values and balancing the groups at 500 (n()-499) observations
```

```
Sub70<-na.omit(bind_cols(combinedHetSNP70, combinedTotSNP70)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs, Individuals = (Individuals/100)*88) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number()>=(n()-999))
```

```
#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals
```

```
AVG70<-Sub70 %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```

HetSNPs = mean(HetSNPs),

TotSNPs = mean(TotSNPs),

PropSNPs = mean(PropSNPs))

#

#Importing All stat observations for different %Individuals at 60% Depth of Coverage

Indiv100_60Depth<-read.csv("100%Indiv_60Depth.csv") %>% mutate(Individuals = rep(100,n())) %>% slice(-
2257:-2646)

Indiv90_60Depth<-read.csv("90%Indiv_60Depth.csv") %>% mutate(Individuals = rep(90,n())) %>% slice(-2257:-
2646)

Indiv80_60Depth<-read.csv("80%Indiv_60Depth.csv") %>% mutate(Individuals = rep(80,n())) %>% slice(-2257:-
2646)

Indiv70_60Depth<-read.csv("70%Indiv_60Depth.csv") %>% mutate(Individuals = rep(70,n())) %>% slice(-2257:-
2646)

Indiv60_60Depth<-read.csv("60%Indiv_60Depth.csv") %>% mutate(Individuals = rep(60,n())) %>% slice(-2257:-
2646)

Indiv50_60Depth<-read.csv("50%Indiv_60Depth.csv") %>% mutate(Individuals = rep(50,n())) %>% slice(-2257:-
2646)

Indiv40_60Depth<-read.csv("40%Indiv_60Depth.csv") %>% mutate(Individuals = rep(40,n())) %>% slice(-2257:-
2646)

Indiv38_60Depth<-read.csv("38%Indiv_60Depth.csv") %>% mutate(Individuals = rep(38,n())) %>% slice(-2257:-
2646)

Indiv36_60Depth<-read.csv("36%Indiv_60Depth.csv") %>% mutate(Individuals = rep(36,n())) %>% slice(-2257:-
2646)

```

```
Indiv34_60Depth<-read.csv("34%Indiv_60Depth.csv") %>% mutate(Individuals = rep(34,n())) %>% slice(-2257:-  
2646)
```

```
Indiv32_60Depth<-read.csv("32%Indiv_60Depth.csv") %>% mutate(Individuals = rep(32,n())) %>% slice(-2257:-  
2646)
```

```
Indiv30_60Depth<-read.csv("30%Indiv_60Depth.csv") %>% mutate(Individuals = rep(30,n())) %>% slice(-2257:-  
2646)
```

```
Indiv28_60Depth<-read.csv("28%Indiv_60Depth.csv") %>% mutate(Individuals = rep(28,n())) %>% slice(-2257:-  
2646)
```

```
Indiv26_60Depth<-read.csv("26%Indiv_60Depth.csv") %>% mutate(Individuals = rep(26,n())) %>% slice(-2257:-  
2646)
```

```
Indiv24_60Depth<-read.csv("24%Indiv_60Depth.csv") %>% mutate(Individuals = rep(24,n())) %>% slice(-2257:-  
2646)
```

```
Indiv22_60Depth<-read.csv("22%Indiv_60Depth.csv") %>% mutate(Individuals = rep(22,n())) %>% slice(-2257:-  
2646)
```

```
Indiv20_60Depth<-read.csv("20%Indiv_60Depth.csv") %>% mutate(Individuals = rep(20,n())) %>% slice(-2257:-  
2646)
```

```
Indiv18_60Depth<-read.csv("18%Indiv_60Depth.csv") %>% mutate(Individuals = rep(18,n())) %>% slice(-2257:-  
2646)
```

```
Indiv16_60Depth<-read.csv("16%Indiv_60Depth.csv") %>% mutate(Individuals = rep(16,n())) %>% slice(-2257:-  
2646)
```

```
Indiv14_60Depth<-read.csv("14%Indiv_60Depth.csv") %>% mutate(Individuals = rep(14,n())) %>% slice(-2257:-  
2646)
```

```
Indiv12_60Depth<-read.csv("12%Indiv_60Depth.csv") %>% mutate(Individuals = rep(12,n())) %>% slice(-2257:-  
2646)
```

```
Indiv10_60Depth<-read.csv("10%Indiv_60Depth.csv") %>% mutate(Individuals = rep(10,n())) %>% slice(-2257:-  
2646)
```

```
Indiv08_60Depth<-read.csv("8%Indiv_60Depth.csv") %>% mutate(Individuals = rep(8,n())) %>% slice(-2257:-  
2646)
```

```
Indiv06_60Depth<-read.csv("6%Indiv_60Depth.csv") %>% mutate(Individuals = rep(6,n())) %>% slice(-2257:-  
2646)
```

```
Indiv04_60Depth<-read.csv("4%Indiv_60Depth.csv") %>% mutate(Individuals = rep(4,n())) %>% slice(-2257:-  
2646)
```

```
Indiv02_60Depth<-read.csv("2%Indiv_60Depth.csv") %>% mutate(Individuals = rep(2,n())) %>% slice(-2257:-  
2646)
```

```
#combine all observations
```

```
combined60<-
```

```
bind_rows(Indiv02_60Depth,Indiv04_60Depth,Indiv06_60Depth,Indiv08_60Depth,Indiv10_60Depth,Indiv12_6  
0Depth,Indiv14_60Depth,Indiv16_60Depth,Indiv18_60Depth,Indiv20_60Depth,Indiv22_60Depth,Indiv24_60D  
epth,Indiv26_60Depth,Indiv28_60Depth,Indiv30_60Depth,Indiv32_60Depth,Indiv34_60Depth,Indiv36_60Dept  
h,Indiv38_60Depth,Indiv40_60Depth,Indiv50_60Depth,Indiv60_60Depth,Indiv70_60Depth,Indiv80_60Depth,In  
div90_60Depth,Indiv100_60Depth)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP60<-combined60 %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value,  
Individuals)
```

```
combinedHetSNP60<-combined60 %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs =  
value, Individuals)
```

#combining these and then creating an additional column containing the proportion of SNPs

```
Sub60<-na.omit(bind_cols(combinedHetSNP60, combinedTotSNP60)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs, Individuals = (Individuals/100)*88) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number()>=(n()-999))
```

#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals

```
AVG60<-Sub60 %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
    HetSNPs = mean(HetSNPs),
```

```
    TotSNPs = mean(TotSNPs),
```

```
    PropSNPs = mean(PropSNPs))
```

#

#Importing All stat observations for different %Individuals at 50% Depth of Coverage

```
Indiv100_50Depth<-read.csv("100%Indiv_50Depth.csv") %>% mutate(Individuals = rep(100,n())) %>% slice(-  
2257:-2646)
```

```
Indiv90_50Depth<-read.csv("90%Indiv_50Depth.csv") %>% mutate(Individuals = rep(90,n())) %>% slice(-2257:-  
2646)
```

```
Indiv80_50Depth<-read.csv("80%Indiv_50Depth.csv") %>% mutate(Individuals = rep(80,n())) %>% slice(-2257:-  
2646)
```

```
Indiv70_50Depth<-read.csv("70%Indiv_50Depth.csv") %>% mutate(Individuals = rep(70,n())) %>% slice(-2257:-  
2646)
```

```
Indiv60_50Depth<-read.csv("60%Indiv_50Depth.csv") %>% mutate(Individuals = rep(60,n())) %>% slice(-2257:-  
2646)
```

```
Indiv50_50Depth<-read.csv("50%Indiv_50Depth.csv") %>% mutate(Individuals = rep(50,n())) %>% slice(-2257:-  
2646)
```

```
Indiv40_50Depth<-read.csv("40%Indiv_50Depth.csv") %>% mutate(Individuals = rep(40,n())) %>% slice(-2257:-  
2646)
```

```
Indiv38_50Depth<-read.csv("38%Indiv_50Depth.csv") %>% mutate(Individuals = rep(38,n())) %>% slice(-2257:-  
2646)
```

```
Indiv36_50Depth<-read.csv("36%Indiv_50Depth.csv") %>% mutate(Individuals = rep(36,n())) %>% slice(-2257:-  
2646)
```

```
Indiv34_50Depth<-read.csv("34%Indiv_50Depth.csv") %>% mutate(Individuals = rep(34,n())) %>% slice(-2257:-  
2646)
```

```
Indiv32_50Depth<-read.csv("32%Indiv_50Depth.csv") %>% mutate(Individuals = rep(32,n())) %>% slice(-2257:-  
2646)
```

```
Indiv30_50Depth<-read.csv("30%Indiv_50Depth.csv") %>% mutate(Individuals = rep(30,n())) %>% slice(-2257:-  
2646)
```

```
Indiv28_50Depth<-read.csv("28%Indiv_50Depth.csv") %>% mutate(Individuals = rep(28,n())) %>% slice(-2257:-  
2646)
```

```
Indiv26_50Depth<-read.csv("26%Indiv_50Depth.csv") %>% mutate(Individuals = rep(26,n())) %>% slice(-2257:-  
2646)
```



```
Indiv24_50Depth<-read.csv("24%Indiv_50Depth.csv") %>% mutate(Individuals = rep(24,n())) %>% slice(-2257:-  
2646)
```

```
Indiv22_50Depth<-read.csv("22%Indiv_50Depth.csv") %>% mutate(Individuals = rep(22,n())) %>% slice(-2257:-  
2646)
```

```
Indiv20_50Depth<-read.csv("20%Indiv_50Depth.csv") %>% mutate(Individuals = rep(20,n())) %>% slice(-2257:-  
2646)
```

```
Indiv18_50Depth<-read.csv("18%Indiv_50Depth.csv") %>% mutate(Individuals = rep(18,n())) %>% slice(-2257:-  
2646)
```

```
Indiv16_50Depth<-read.csv("16%Indiv_50Depth.csv") %>% mutate(Individuals = rep(16,n())) %>% slice(-2257:-  
2646)
```

```
Indiv14_50Depth<-read.csv("14%Indiv_50Depth.csv") %>% mutate(Individuals = rep(14,n())) %>% slice(-2257:-  
2646)
```

```
Indiv12_50Depth<-read.csv("12%Indiv_50Depth.csv") %>% mutate(Individuals = rep(12,n())) %>% slice(-2257:-  
2646)
```

```
Indiv10_50Depth<-read.csv("10%Indiv_50Depth.csv") %>% mutate(Individuals = rep(10,n())) %>% slice(-2257:-  
2646)
```

```
Indiv08_50Depth<-read.csv("8%Indiv_50Depth.csv") %>% mutate(Individuals = rep(8,n())) %>% slice(-2257:-  
2646)
```

```
Indiv06_50Depth<-read.csv("6%Indiv_50Depth.csv") %>% mutate(Individuals = rep(6,n())) %>% slice(-2257:-  
2646)
```

```
Indiv04_50Depth<-read.csv("4%Indiv_50Depth.csv") %>% mutate(Individuals = rep(4,n())) %>% slice(-2257:-  
2646)
```

```
Indiv02_50Depth<-read.csv("2%Indiv_50Depth.csv") %>% mutate(Individuals = rep(2,n())) %>% slice(-2257:-  
2646)
```

```
#combine all observations
```

```
combined50<-
```

```
bind_rows(Indiv02_50Depth,Indiv04_50Depth,Indiv06_50Depth,Indiv08_50Depth,Indiv10_50Depth,Indiv12_50Depth,Indiv14_50Depth,Indiv16_50Depth,Indiv18_50Depth,Indiv20_50Depth,Indiv22_50Depth,Indiv24_50Depth,Indiv26_50Depth,Indiv28_50Depth,Indiv30_50Depth,Indiv32_50Depth,Indiv34_50Depth,Indiv36_50Depth,Indiv38_50Depth,Indiv40_50Depth,Indiv50_50Depth,Indiv60_50Depth,Indiv70_50Depth,Indiv80_50Depth,Indiv90_50Depth,Indiv100_50Depth)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP50<-combined50 %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
combinedHetSNP50<-combined50 %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```

```
Sub50<-na.omit(bind_cols(combinedHetSNP50, combinedTotSNP50)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs, Individuals = (Individuals/100)*88) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number()>=(n()-999))
```

```
#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals
```

```
AVG50<-Sub50 %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
    HetSNPs = mean(HetSNPs),
```

```
    TotSNPs = mean(TotSNPs),
```

```
    PropSNPs = mean(PropSNPs))
```

```
#
```

```
#Importing All stat observations for different %Individuals at 40% Depth of Coverage
```

```
Indiv100_40Depth<-read.csv("100%Indiv_40Depth.csv") %>% mutate(Individuals = rep(100,n())) %>% slice(-  
2257:-2646)
```

```
Indiv90_40Depth<-read.csv("90%Indiv_40Depth.csv") %>% mutate(Individuals = rep(90,n())) %>% slice(-2257:-  
2646)
```

```
Indiv80_40Depth<-read.csv("80%Indiv_40Depth.csv") %>% mutate(Individuals = rep(80,n())) %>% slice(-2257:-  
2646)
```

```
Indiv70_40Depth<-read.csv("70%Indiv_40Depth.csv") %>% mutate(Individuals = rep(70,n())) %>% slice(-2257:-  
2646)
```

```
Indiv60_40Depth<-read.csv("60%Indiv_40Depth.csv") %>% mutate(Individuals = rep(60,n())) %>% slice(-2257:-  
2646)
```

```
Indiv50_40Depth<-read.csv("50%Indiv_40Depth.csv") %>% mutate(Individuals = rep(50,n())) %>% slice(-2257:-  
2646)
```

```
Indiv40_40Depth<-read.csv("40%Indiv_40Depth.csv") %>% mutate(Individuals = rep(40,n())) %>% slice(-2257:-  
2646)
```

```
Indiv38_40Depth<-read.csv("38%Indiv_40Depth.csv") %>% mutate(Individuals = rep(38,n())) %>% slice(-2257:-  
2646)
```

```
Indiv36_40Depth<-read.csv("36%Indiv_40Depth.csv") %>% mutate(Individuals = rep(36,n())) %>% slice(-2257:-  
2646)
```

```
Indiv34_40Depth<-read.csv("34%Indiv_40Depth.csv") %>% mutate(Individuals = rep(34,n())) %>% slice(-2257:-  
2646)
```

```
Indiv32_40Depth<-read.csv("32%Indiv_40Depth.csv") %>% mutate(Individuals = rep(32,n())) %>% slice(-2257:-  
2646)
```

```
Indiv30_40Depth<-read.csv("30%Indiv_40Depth.csv") %>% mutate(Individuals = rep(30,n())) %>% slice(-2257:-  
2646)
```

```
Indiv28_40Depth<-read.csv("28%Indiv_40Depth.csv") %>% mutate(Individuals = rep(28,n())) %>% slice(-2257:-  
2646)
```

```
Indiv26_40Depth<-read.csv("26%Indiv_40Depth.csv") %>% mutate(Individuals = rep(26,n())) %>% slice(-2257:-  
2646)
```

```
Indiv24_40Depth<-read.csv("24%Indiv_40Depth.csv") %>% mutate(Individuals = rep(24,n())) %>% slice(-2257:-  
2646)
```

```
Indiv22_40Depth<-read.csv("22%Indiv_40Depth.csv") %>% mutate(Individuals = rep(22,n())) %>% slice(-2257:-  
2646)
```

```
Indiv20_40Depth<-read.csv("20%Indiv_40Depth.csv") %>% mutate(Individuals = rep(20,n())) %>% slice(-2257:-  
2646)
```

```
Indiv18_40Depth<-read.csv("18%Indiv_40Depth.csv") %>% mutate(Individuals = rep(18,n())) %>% slice(-2257:-  
2646)
```

```
Indiv16_40Depth<-read.csv("16%Indiv_40Depth.csv") %>% mutate(Individuals = rep(16,n())) %>% slice(-2257:-  
2646)
```

```
Indiv14_40Depth<-read.csv("14%Indiv_40Depth.csv") %>% mutate(Individuals = rep(14,n())) %>% slice(-2257:-2646)
```

```
Indiv12_40Depth<-read.csv("12%Indiv_40Depth.csv") %>% mutate(Individuals = rep(12,n())) %>% slice(-2257:-2646)
```

```
Indiv10_40Depth<-read.csv("10%Indiv_40Depth.csv") %>% mutate(Individuals = rep(10,n())) %>% slice(-2257:-2646)
```

```
Indiv08_40Depth<-read.csv("8%Indiv_40Depth.csv") %>% mutate(Individuals = rep(8,n())) %>% slice(-2257:-2646)
```

```
Indiv06_40Depth<-read.csv("6%Indiv_40Depth.csv") %>% mutate(Individuals = rep(6,n())) %>% slice(-2257:-2646)
```

```
Indiv04_40Depth<-read.csv("4%Indiv_40Depth.csv") %>% mutate(Individuals = rep(4,n())) %>% slice(-2257:-2646)
```

```
Indiv02_40Depth<-read.csv("2%Indiv_40Depth.csv") %>% mutate(Individuals = rep(2,n())) %>% slice(-2257:-2646)
```

```
#combine all observations
```

```
combined40<-
```

```
bind_rows(Indiv02_40Depth,Indiv04_40Depth,Indiv06_40Depth,Indiv08_40Depth,Indiv10_40Depth,Indiv12_40Depth,Indiv14_40Depth,Indiv16_40Depth,Indiv18_40Depth,Indiv20_40Depth,Indiv22_40Depth,Indiv24_40Depth,Indiv26_40Depth,Indiv28_40Depth,Indiv30_40Depth,Indiv32_40Depth,Indiv34_40Depth,Indiv36_40Depth,Indiv38_40Depth,Indiv40_40Depth,Indiv50_40Depth,Indiv60_40Depth,Indiv70_40Depth,Indiv80_40Depth,Indiv90_40Depth,Indiv100_40Depth)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP40<-combined40 %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value,  
Individuals)
```

```
combinedHetSNP40<-combined40 %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs =  
value, Individuals)
```

#combining these and then creating an additional column containing the proportion of SNPs

```
Sub40<-na.omit(bind_cols(combinedHetSNP40, combinedTotSNP40)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs, Individuals = (Individuals/100)*88) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number())>=(n()-999))
```

#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals

```
AVG40<-Sub40 %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
  HetSNPs = mean(HetSNPs),
```

```
  TotSNPs = mean(TotSNPs),
```

```
  PropSNPs = mean(PropSNPs))
```

```
#
```

#Importing All stat observations for different %Individuals at 30% Depth of Coverage

```
Indiv100_30Depth<-read.csv("100%Indiv_30Depth.csv") %>% mutate(Individuals = rep(100,n())) %>% slice(-  
2257:-2646)
```

```
Indiv90_30Depth<-read.csv("90%Indiv_30Depth.csv") %>% mutate(Individuals = rep(90,n())) %>% slice(-2257:-  
2646)
```

```
Indiv80_30Depth<-read.csv("80%Indiv_30Depth.csv") %>% mutate(Individuals = rep(80,n())) %>% slice(-2257:-  
2646)
```

```
Indiv70_30Depth<-read.csv("70%Indiv_30Depth.csv") %>% mutate(Individuals = rep(70,n())) %>% slice(-2257:-  
2646)
```

```
Indiv60_30Depth<-read.csv("60%Indiv_30Depth.csv") %>% mutate(Individuals = rep(60,n())) %>% slice(-2257:-  
2646)
```

```
Indiv50_30Depth<-read.csv("50%Indiv_30Depth.csv") %>% mutate(Individuals = rep(50,n())) %>% slice(-2257:-  
2646)
```

```
Indiv40_30Depth<-read.csv("40%Indiv_30Depth.csv") %>% mutate(Individuals = rep(40,n())) %>% slice(-2257:-  
2646)
```

```
Indiv38_30Depth<-read.csv("38%Indiv_30Depth.csv") %>% mutate(Individuals = rep(38,n())) %>% slice(-2257:-  
2646)
```

```
Indiv36_30Depth<-read.csv("36%Indiv_30Depth.csv") %>% mutate(Individuals = rep(36,n())) %>% slice(-2257:-  
2646)
```

```
Indiv34_30Depth<-read.csv("34%Indiv_30Depth.csv") %>% mutate(Individuals = rep(34,n())) %>% slice(-2257:-  
2646)
```

```
Indiv32_30Depth<-read.csv("32%Indiv_30Depth.csv") %>% mutate(Individuals = rep(32,n())) %>% slice(-2257:-  
2646)
```

```
Indiv30_30Depth<-read.csv("30%Indiv_30Depth.csv") %>% mutate(Individuals = rep(30,n())) %>% slice(-2257:-  
2646)
```

```
Indiv28_30Depth<-read.csv("28%Indiv_30Depth.csv") %>% mutate(Individuals = rep(28,n())) %>% slice(-2257:-  
2646)
```

```
Indiv26_30Depth<-read.csv("26%Indiv_30Depth.csv") %>% mutate(Individuals = rep(26,n())) %>% slice(-2257:-  
2646)
```

```
Indiv24_30Depth<-read.csv("24%Indiv_30Depth.csv") %>% mutate(Individuals = rep(24,n())) %>% slice(-2257:-  
2646)
```

```
Indiv22_30Depth<-read.csv("22%Indiv_30Depth.csv") %>% mutate(Individuals = rep(22,n())) %>% slice(-2257:-  
2646)
```

```
Indiv20_30Depth<-read.csv("20%Indiv_30Depth.csv") %>% mutate(Individuals = rep(20,n())) %>% slice(-2257:-  
2646)
```

```
Indiv18_30Depth<-read.csv("18%Indiv_30Depth.csv") %>% mutate(Individuals = rep(18,n())) %>% slice(-2257:-  
2646)
```

```
Indiv16_30Depth<-read.csv("16%Indiv_30Depth.csv") %>% mutate(Individuals = rep(16,n())) %>% slice(-2257:-  
2646)
```

```
Indiv14_30Depth<-read.csv("14%Indiv_30Depth.csv") %>% mutate(Individuals = rep(14,n())) %>% slice(-2257:-  
2646)
```

```
Indiv12_30Depth<-read.csv("12%Indiv_30Depth.csv") %>% mutate(Individuals = rep(12,n())) %>% slice(-2257:-  
2646)
```

```
Indiv10_30Depth<-read.csv("10%Indiv_30Depth.csv") %>% mutate(Individuals = rep(10,n())) %>% slice(-2257:-  
2646)
```

```
Indiv08_30Depth<-read.csv("8%Indiv_30Depth.csv") %>% mutate(Individuals = rep(8,n())) %>% slice(-2257:-  
2646)
```

```
Indiv06_30Depth<-read.csv("6%Indiv_30Depth.csv") %>% mutate(Individuals = rep(6,n())) %>% slice(-2257:-  
2646)
```



```
Indiv04_30Depth<-read.csv("4%Indiv_30Depth.csv") %>% mutate(Individuals = rep(4,n())) %>% slice(-2257:-2646)
```

```
Indiv02_30Depth<-read.csv("2%Indiv_30Depth.csv") %>% mutate(Individuals = rep(2,n())) %>% slice(-2257:-2646)
```

```
#combine all observations
```

```
combined30<-
```

```
bind_rows(Indiv02_30Depth,Indiv04_30Depth,Indiv06_30Depth,Indiv08_30Depth,Indiv10_30Depth,Indiv12_30Depth,Indiv14_30Depth,Indiv16_30Depth,Indiv18_30Depth,Indiv20_30Depth,Indiv22_30Depth,Indiv24_30Depth,Indiv26_30Depth,Indiv28_30Depth,Indiv30_30Depth,Indiv32_30Depth,Indiv34_30Depth,Indiv36_30Depth,Indiv38_30Depth,Indiv40_30Depth,Indiv50_30Depth,Indiv60_30Depth,Indiv70_30Depth,Indiv80_30Depth,Indiv90_30Depth,Indiv100_30Depth)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP30<-combined30 %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
combinedHetSNP30<-combined30 %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```

```
Sub30<-na.omit(bind_cols(combinedHetSNP30, combinedTotSNP30)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs, Individuals = (Individuals/100)*88) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number())>=(n()-999))
```

```
#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals
```

```
AVG30<-Sub30 %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
    HetSNPs = mean(HetSNPs),
```

```
    TotSNPs = mean(TotSNPs),
```

```
    PropSNPs = mean(PropSNPs))
```

```
#
```

```
#Importing All stat observations for different %Individuals at 20% Depth of Coverage
```

```
Indiv100_20Depth<-read.csv("100%Indiv_20Depth.csv") %>% mutate(Individuals = rep(100,n())) %>% slice(-2257:-2646)
```

```
Indiv90_20Depth<-read.csv("90%Indiv_20Depth.csv") %>% mutate(Individuals = rep(90,n())) %>% slice(-2257:-2646)
```

```
Indiv80_20Depth<-read.csv("80%Indiv_20Depth.csv") %>% mutate(Individuals = rep(80,n())) %>% slice(-2257:-2646)
```

```
Indiv70_20Depth<-read.csv("70%Indiv_20Depth.csv") %>% mutate(Individuals = rep(70,n())) %>% slice(-2257:-2646)
```

```
Indiv60_20Depth<-read.csv("60%Indiv_20Depth.csv") %>% mutate(Individuals = rep(60,n())) %>% slice(-2257:-2646)
```

```
Indiv50_20Depth<-read.csv("50%Indiv_20Depth.csv") %>% mutate(Individuals = rep(50,n())) %>% slice(-2257:-  
2646)
```

```
Indiv40_20Depth<-read.csv("40%Indiv_20Depth.csv") %>% mutate(Individuals = rep(40,n())) %>% slice(-2257:-  
2646)
```

```
Indiv38_20Depth<-read.csv("38%Indiv_20Depth.csv") %>% mutate(Individuals = rep(38,n())) %>% slice(-2257:-  
2646)
```

```
Indiv36_20Depth<-read.csv("36%Indiv_20Depth.csv") %>% mutate(Individuals = rep(36,n())) %>% slice(-2257:-  
2646)
```

```
Indiv34_20Depth<-read.csv("34%Indiv_20Depth.csv") %>% mutate(Individuals = rep(34,n())) %>% slice(-2257:-  
2646)
```

```
Indiv32_20Depth<-read.csv("32%Indiv_20Depth.csv") %>% mutate(Individuals = rep(32,n())) %>% slice(-2257:-  
2646)
```

```
Indiv30_20Depth<-read.csv("30%Indiv_20Depth.csv") %>% mutate(Individuals = rep(30,n())) %>% slice(-2257:-  
2646)
```

```
Indiv28_20Depth<-read.csv("28%Indiv_20Depth.csv") %>% mutate(Individuals = rep(28,n())) %>% slice(-2257:-  
2646)
```

```
Indiv26_20Depth<-read.csv("26%Indiv_20Depth.csv") %>% mutate(Individuals = rep(26,n())) %>% slice(-2257:-  
2646)
```

```
Indiv24_20Depth<-read.csv("24%Indiv_20Depth.csv") %>% mutate(Individuals = rep(24,n())) %>% slice(-2257:-  
2646)
```

```
Indiv22_20Depth<-read.csv("22%Indiv_20Depth.csv") %>% mutate(Individuals = rep(22,n())) %>% slice(-2257:-  
2646)
```

```
Indiv20_20Depth<-read.csv("20%Indiv_20Depth.csv") %>% mutate(Individuals = rep(20,n())) %>% slice(-2257:-  
2646)
```

```
Indiv18_20Depth<-read.csv("18%Indiv_20Depth.csv") %>% mutate(Individuals = rep(18,n())) %>% slice(-2257:-  
2646)
```

```
Indiv16_20Depth<-read.csv("16%Indiv_20Depth.csv") %>% mutate(Individuals = rep(16,n())) %>% slice(-2257:-  
2646)
```

```
Indiv14_20Depth<-read.csv("14%Indiv_20Depth.csv") %>% mutate(Individuals = rep(14,n())) %>% slice(-2257:-  
2646)
```

```
Indiv12_20Depth<-read.csv("12%Indiv_20Depth.csv") %>% mutate(Individuals = rep(12,n())) %>% slice(-2257:-  
2646)
```

```
Indiv10_20Depth<-read.csv("10%Indiv_20Depth.csv") %>% mutate(Individuals = rep(10,n())) %>% slice(-2257:-  
2646)
```

```
Indiv08_20Depth<-read.csv("8%Indiv_20Depth.csv") %>% mutate(Individuals = rep(8,n())) %>% slice(-2257:-  
2646)
```

```
Indiv06_20Depth<-read.csv("6%Indiv_20Depth.csv") %>% mutate(Individuals = rep(6,n())) %>% slice(-2257:-  
2646)
```

```
Indiv04_20Depth<-read.csv("4%Indiv_20Depth.csv") %>% mutate(Individuals = rep(4,n())) %>% slice(-2257:-  
2646)
```

```
Indiv02_20Depth<-read.csv("2%Indiv_20Depth.csv") %>% mutate(Individuals = rep(2,n())) %>% slice(-2257:-  
2646)
```

```
#combine all observations
```

```
combined20<-
```

```
bind_rows(Indiv02_20Depth,Indiv04_20Depth,Indiv06_20Depth,Indiv08_20Depth,Indiv10_20Depth,Indiv12_2  
0Depth,Indiv14_20Depth,Indiv16_20Depth,Indiv18_20Depth,Indiv20_20Depth,Indiv22_20Depth,Indiv24_20D  
epth,Indiv26_20Depth,Indiv28_20Depth,Indiv30_20Depth,Indiv32_20Depth,Indiv34_20Depth,Indiv36_20Dept
```

```
h,Indiv38_20Depth,Indiv40_20Depth,Indiv50_20Depth,Indiv60_20Depth,Indiv70_20Depth,Indiv80_20Depth,Indiv90_20Depth,Indiv100_20Depth)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP20<-combined20 %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
combinedHetSNP20<-combined20 %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```

```
Sub20<-na.omit(bind_cols(combinedHetSNP20, combinedTotSNP20)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs, Individuals = (Individuals/100)*88) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number())>=(n()-999))
```

```
#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals
```

```
AVG20<-Sub20 %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
  HetSNPs = mean(HetSNPs),
```

```
  TotSNPs = mean(TotSNPs),
```

```
PropSNPs = mean(PropSNPs))
```

```
#
```

```
#Importing All stat observations for different %Individuals at 10% Depth of Coverage
```

```
Indiv100_10Depth<-read.csv("100%Indiv_10Depth.csv") %>% mutate(Individuals = rep(100,n())) %>% slice(-  
2257:-2646)
```

```
Indiv90_10Depth<-read.csv("90%Indiv_10Depth.csv") %>% mutate(Individuals = rep(90,n())) %>% slice(-2257:-  
2646)
```

```
Indiv80_10Depth<-read.csv("80%Indiv_10Depth.csv") %>% mutate(Individuals = rep(80,n())) %>% slice(-2257:-  
2646)
```

```
Indiv70_10Depth<-read.csv("70%Indiv_10Depth.csv") %>% mutate(Individuals = rep(70,n())) %>% slice(-2257:-  
2646)
```

```
Indiv60_10Depth<-read.csv("60%Indiv_10Depth.csv") %>% mutate(Individuals = rep(60,n())) %>% slice(-2257:-  
2646)
```

```
Indiv50_10Depth<-read.csv("50%Indiv_10Depth.csv") %>% mutate(Individuals = rep(50,n())) %>% slice(-2257:-  
2646)
```

```
Indiv40_10Depth<-read.csv("40%Indiv_10Depth.csv") %>% mutate(Individuals = rep(40,n())) %>% slice(-2257:-  
2646)
```

```
Indiv38_10Depth<-read.csv("38%Indiv_10Depth.csv") %>% mutate(Individuals = rep(38,n())) %>% slice(-2257:-  
2646)
```

```
Indiv36_10Depth<-read.csv("36%Indiv_10Depth.csv") %>% mutate(Individuals = rep(36,n())) %>% slice(-2257:-  
2646)
```

```
Indiv34_10Depth<-read.csv("34%Indiv_10Depth.csv") %>% mutate(Individuals = rep(34,n())) %>% slice(-2257:-  
2646)
```

```
Indiv32_10Depth<-read.csv("32%Indiv_10Depth.csv") %>% mutate(Individuals = rep(32,n())) %>% slice(-2257:-  
2646)
```

```
Indiv30_10Depth<-read.csv("30%Indiv_10Depth.csv") %>% mutate(Individuals = rep(30,n())) %>% slice(-2257:-  
2646)
```

```
Indiv28_10Depth<-read.csv("28%Indiv_10Depth.csv") %>% mutate(Individuals = rep(28,n())) %>% slice(-2257:-  
2646)
```

```
Indiv26_10Depth<-read.csv("26%Indiv_10Depth.csv") %>% mutate(Individuals = rep(26,n())) %>% slice(-2257:-  
2646)
```

```
Indiv24_10Depth<-read.csv("24%Indiv_10Depth.csv") %>% mutate(Individuals = rep(24,n())) %>% slice(-2257:-  
2646)
```

```
Indiv22_10Depth<-read.csv("22%Indiv_10Depth.csv") %>% mutate(Individuals = rep(22,n())) %>% slice(-2257:-  
2646)
```

```
Indiv20_10Depth<-read.csv("20%Indiv_10Depth.csv") %>% mutate(Individuals = rep(20,n())) %>% slice(-2257:-  
2646)
```

```
Indiv18_10Depth<-read.csv("18%Indiv_10Depth.csv") %>% mutate(Individuals = rep(18,n())) %>% slice(-2257:-  
2646)
```

```
Indiv16_10Depth<-read.csv("16%Indiv_10Depth.csv") %>% mutate(Individuals = rep(16,n())) %>% slice(-2257:-  
2646)
```

```
Indiv14_10Depth<-read.csv("14%Indiv_10Depth.csv") %>% mutate(Individuals = rep(14,n())) %>% slice(-2257:-  
2646)
```

```
Indiv12_10Depth<-read.csv("12%Indiv_10Depth.csv") %>% mutate(Individuals = rep(12,n())) %>% slice(-2257:-  
2646)
```

```
Indiv10_10Depth<-read.csv("10%Indiv_10Depth.csv") %>% mutate(Individuals = rep(10,n())) %>% slice(-2257:-  
2646)
```

```
Indiv08_10Depth<-read.csv("8%Indiv_10Depth.csv") %>% mutate(Individuals = rep(8,n())) %>% slice(-2257:-2646)
```

```
Indiv06_10Depth<-read.csv("6%Indiv_10Depth.csv") %>% mutate(Individuals = rep(6,n())) %>% slice(-2257:-2646)
```

```
Indiv04_10Depth<-read.csv("4%Indiv_10Depth.csv") %>% mutate(Individuals = rep(4,n())) %>% slice(-2257:-2646)
```

```
Indiv02_10Depth<-read.csv("2%Indiv_10Depth.csv") %>% mutate(Individuals = rep(2,n())) %>% slice(-2257:-2646)
```

```
#combine all observations
```

```
combined10<-
```

```
bind_rows(Indiv02_10Depth,Indiv04_10Depth,Indiv06_10Depth,Indiv08_10Depth,Indiv10_10Depth,Indiv12_10Depth,Indiv14_10Depth,Indiv16_10Depth,Indiv18_10Depth,Indiv20_10Depth,Indiv22_10Depth,Indiv24_10Depth,Indiv26_10Depth,Indiv28_10Depth,Indiv30_10Depth,Indiv32_10Depth,Indiv34_10Depth,Indiv36_10Depth,Indiv38_10Depth,Indiv40_10Depth,Indiv50_10Depth,Indiv60_10Depth,Indiv70_10Depth,Indiv80_10Depth,Indiv90_10Depth,Indiv100_10Depth)
```

```
#creating subsets with each of the relevant statistics
```

```
combinedTotSNP10<-combined10 %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
combinedHetSNP10<-combined10 %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```



```
Sub10<-na.omit(bind_cols(combinedHetSNP10, combinedTotSNP10)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs, Individuals = (Individuals/100)*88) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number())>=(n()-999))
```

#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals

```
AVG10<-Sub10 %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
  HetSNPs = mean(HetSNPs),
```

```
  TotSNPs = mean(TotSNPs),
```

```
  PropSNPs = mean(PropSNPs))
```

[R script for Buller's Albatross](#)

```
library(dplyr)
```

```
setwd("/media/levi/1TB/Bullers/Subsampled_VCFs/csvs_with_headers")
```

```
#Importing All stat observations for different Individuals for the combined bullers dataset
```

```
Indiv95_bullers_all<-read.csv("95Indiv_bullers_all.csv") %>% mutate(Individuals = rep(95,n()))
```

```
Indiv90_bullers_all<-read.csv("90Indiv_bullers_all.csv") %>% mutate(Individuals = rep(90,n()))

Indiv80_bullers_all<-read.csv("80Indiv_bullers_all.csv") %>% mutate(Individuals = rep(80,n()))

Indiv70_bullers_all<-read.csv("70Indiv_bullers_all.csv") %>% mutate(Individuals = rep(70,n()))

Indiv60_bullers_all<-read.csv("60Indiv_bullers_all.csv") %>% mutate(Individuals = rep(60,n()))

Indiv50_bullers_all<-read.csv("50Indiv_bullers_all.csv") %>% mutate(Individuals = rep(50,n()))

Indiv40_bullers_all<-read.csv("40Indiv_bullers_all.csv") %>% mutate(Individuals = rep(40,n()))

Indiv38_bullers_all<-read.csv("38Indiv_bullers_all.csv") %>% mutate(Individuals = rep(38,n()))

Indiv36_bullers_all<-read.csv("36Indiv_bullers_all.csv") %>% mutate(Individuals = rep(36,n()))

Indiv34_bullers_all<-read.csv("34Indiv_bullers_all.csv") %>% mutate(Individuals = rep(34,n()))

Indiv32_bullers_all<-read.csv("32Indiv_bullers_all.csv") %>% mutate(Individuals = rep(32,n()))

Indiv30_bullers_all<-read.csv("30Indiv_bullers_all.csv") %>% mutate(Individuals = rep(30,n()))

Indiv28_bullers_all<-read.csv("28Indiv_bullers_all.csv") %>% mutate(Individuals = rep(28,n()))

Indiv26_bullers_all<-read.csv("26Indiv_bullers_all.csv") %>% mutate(Individuals = rep(26,n()))

Indiv24_bullers_all<-read.csv("24Indiv_bullers_all.csv") %>% mutate(Individuals = rep(24,n()))

Indiv22_bullers_all<-read.csv("22Indiv_bullers_all.csv") %>% mutate(Individuals = rep(22,n()))

Indiv20_bullers_all<-read.csv("20Indiv_bullers_all.csv") %>% mutate(Individuals = rep(20,n()))

Indiv18_bullers_all<-read.csv("18Indiv_bullers_all.csv") %>% mutate(Individuals = rep(18,n()))

Indiv16_bullers_all<-read.csv("16Indiv_bullers_all.csv") %>% mutate(Individuals = rep(16,n()))

Indiv14_bullers_all<-read.csv("14Indiv_bullers_all.csv") %>% mutate(Individuals = rep(14,n()))

Indiv12_bullers_all<-read.csv("12Indiv_bullers_all.csv") %>% mutate(Individuals = rep(12,n()))

Indiv10_bullers_all<-read.csv("10Indiv_bullers_all.csv") %>% mutate(Individuals = rep(10,n()))

Indiv08_bullers_all<-read.csv("8Indiv_bullers_all.csv") %>% mutate(Individuals = rep(8,n()))
```

```
Indiv06_bullers_all<-read.csv("6Indiv_bullers_all.csv") %>% mutate(Individuals = rep(6,n()))
```

```
Indiv04_bullers_all<-read.csv("4Indiv_bullers_all.csv") %>% mutate(Individuals = rep(4,n()))
```

```
Indiv02_bullers_all<-read.csv("2Indiv_bullers_all.csv") %>% mutate(Individuals = rep(2,n()))
```

```
#combine all observations
```

```
bullers_allcombined<-
```

```
bind_rows(Indiv02_bullers_all,Indiv04_bullers_all,Indiv06_bullers_all,Indiv08_bullers_all,Indiv10_bullers_all,Indiv12_bullers_all,Indiv14_bullers_all,Indiv16_bullers_all,Indiv18_bullers_all,Indiv20_bullers_all,Indiv22_bullers_all,Indiv24_bullers_all,Indiv26_bullers_all,Indiv28_bullers_all,Indiv30_bullers_all,Indiv32_bullers_all,Indiv34_bullers_all,Indiv36_bullers_all,Indiv38_bullers_all,Indiv40_bullers_all,Indiv50_bullers_all,Indiv60_bullers_all,Indiv70_bullers_all,Indiv80_bullers_all,Indiv90_bullers_all,Indiv95_bullers_all)
```

```
#creating bullers_allSubsets with each of the relevant statistics
```

```
bullers_allcombinedTotSNP<-bullers_allcombined %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
bullers_allcombinedHetSNP<-bullers_allcombined %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
bullers_allcombinedHetENT<-bullers_allcombined %>% filter(stat == "Total Heterozygous entries min. 3") %>% select(HetENTs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```

```
bullers_allSub<-na.omit(bind_cols(bullers_allcombinedHetSNP, bullers_allcombinedTotSNP,  
bullers_allcombinedHetENT)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs, HetENTs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number())>=(n()-999))
```

#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals (because some observations are lost as NA values during data generation)

```
bullers_allAVG<-bullers_allSub %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
  HetSNPs = mean(HetSNPs),
```

```
  TotSNPs = mean(TotSNPs),
```

```
  PropSNPs = mean(PropSNPs),
```

```
  HetENTs = mean(HetENTs))
```

```
#
```

```
#Importing All stat observations for Southern bullers population
```

```
Indiv67_bullers_S<-read.csv("67Indiv_bullers_S.csv") %>% mutate(Individuals = rep(67,n()))
```

```
Indiv60_bullers_S<-read.csv("60Indiv_bullers_S.csv") %>% mutate(Individuals = rep(60,n()))

Indiv50_bullers_S<-read.csv("50Indiv_bullers_S.csv") %>% mutate(Individuals = rep(50,n()))

Indiv40_bullers_S<-read.csv("40Indiv_bullers_S.csv") %>% mutate(Individuals = rep(40,n()))

Indiv38_bullers_S<-read.csv("38Indiv_bullers_S.csv") %>% mutate(Individuals = rep(38,n()))

Indiv36_bullers_S<-read.csv("36Indiv_bullers_S.csv") %>% mutate(Individuals = rep(36,n()))

Indiv34_bullers_S<-read.csv("34Indiv_bullers_S.csv") %>% mutate(Individuals = rep(34,n()))

Indiv32_bullers_S<-read.csv("32Indiv_bullers_S.csv") %>% mutate(Individuals = rep(32,n()))

Indiv30_bullers_S<-read.csv("30Indiv_bullers_S.csv") %>% mutate(Individuals = rep(30,n()))

Indiv28_bullers_S<-read.csv("28Indiv_bullers_S.csv") %>% mutate(Individuals = rep(28,n()))

Indiv26_bullers_S<-read.csv("26Indiv_bullers_S.csv") %>% mutate(Individuals = rep(26,n()))

Indiv24_bullers_S<-read.csv("24Indiv_bullers_S.csv") %>% mutate(Individuals = rep(24,n()))

Indiv22_bullers_S<-read.csv("22Indiv_bullers_S.csv") %>% mutate(Individuals = rep(22,n()))

Indiv20_bullers_S<-read.csv("20Indiv_bullers_S.csv") %>% mutate(Individuals = rep(20,n()))

Indiv18_bullers_S<-read.csv("18Indiv_bullers_S.csv") %>% mutate(Individuals = rep(18,n()))

Indiv16_bullers_S<-read.csv("16Indiv_bullers_S.csv") %>% mutate(Individuals = rep(16,n()))

Indiv14_bullers_S<-read.csv("14Indiv_bullers_S.csv") %>% mutate(Individuals = rep(14,n()))

Indiv12_bullers_S<-read.csv("12Indiv_bullers_S.csv") %>% mutate(Individuals = rep(12,n()))

Indiv10_bullers_S<-read.csv("10Indiv_bullers_S.csv") %>% mutate(Individuals = rep(10,n()))

Indiv08_bullers_S<-read.csv("8Indiv_bullers_S.csv") %>% mutate(Individuals = rep(8,n()))

Indiv06_bullers_S<-read.csv("6Indiv_bullers_S.csv") %>% mutate(Individuals = rep(6,n()))

Indiv04_bullers_S<-read.csv("4Indiv_bullers_S.csv") %>% mutate(Individuals = rep(4,n()))

Indiv02_bullers_S<-read.csv("2Indiv_bullers_S.csv") %>% mutate(Individuals = rep(2,n()))
```

```
#combine all observations
```

```
bullers_Scombined<-
```

```
bind_rows(Indiv02_bullers_S,Indiv04_bullers_S,Indiv06_bullers_S,Indiv08_bullers_S,Indiv10_bullers_S,Indiv12_bullers_S,Indiv14_bullers_S,Indiv16_bullers_S,Indiv18_bullers_S,Indiv20_bullers_S,Indiv22_bullers_S,Indiv24_bullers_S,Indiv26_bullers_S,Indiv28_bullers_S,Indiv30_bullers_S,Indiv32_bullers_S,Indiv34_bullers_S,Indiv36_bullers_S,Indiv38_bullers_S,Indiv40_bullers_S,Indiv50_bullers_S,Indiv60_bullers_S,Indiv67_bullers_S)
```

```
#creating bullers_SSubsets with each of the relevant statistics
```

```
bullers_ScombinedTotSNP<-bullers_Scombined %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
bullers_ScombinedHetSNP<-bullers_Scombined %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
bullers_ScombinedHetENT<-bullers_Scombined %>% filter(stat == "Total Heterozygous entries min. 3") %>% select(HetENTs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```

```
bullers_SSub<-na.omit(bind_cols(bullers_ScombinedHetSNP, bullers_ScombinedTotSNP, bullers_ScombinedHetENT)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs, HetENTs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs) %>%
```

```
group_by(Individuals) %>%
```

```
filter(row_number()>=(n()-999))
```

#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals (because some observations are lost as NA values during data generation)

```
bullers_SAVG<-bullers_SSub %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
  HetSNPs = mean(HetSNPs),
```

```
  TotSNPs = mean(TotSNPs),
```

```
  PropSNPs = mean(PropSNPs),
```

```
  HetENTs = mean(HetENTs))
```

#Importing All stat observations for Northern bullers population

```
Indiv28_bullers_N<-read.csv("28Indiv_bullers_N.csv") %>% mutate(Individuals = rep(28,n()))
```

```
Indiv26_bullers_N<-read.csv("26Indiv_bullers_N.csv") %>% mutate(Individuals = rep(26,n()))
```

```
Indiv24_bullers_N<-read.csv("24Indiv_bullers_N.csv") %>% mutate(Individuals = rep(24,n()))
```

```
Indiv22_bullers_N<-read.csv("22Indiv_bullers_N.csv") %>% mutate(Individuals = rep(22,n()))
```

```
Indiv20_bullers_N<-read.csv("20Indiv_bullers_N.csv") %>% mutate(Individuals = rep(20,n()))
```

```
Indiv18_bullers_N<-read.csv("18Indiv_bullers_N.csv") %>% mutate(Individuals = rep(18,n()))
```

```
Indiv16_bullers_N<-read.csv("16Indiv_bullers_N.csv") %>% mutate(Individuals = rep(16,n()))
```

```
Indiv14_bullers_N<-read.csv("14Indiv_bullers_N.csv") %>% mutate(Individuals = rep(14,n()))
```

```
Indiv12_bullers_N<-read.csv("12Indiv_bullers_N.csv") %>% mutate(Individuals = rep(12,n()))
```

```
Indiv10_bullers_N<-read.csv("10Indiv_bullers_N.csv") %>% mutate(Individuals = rep(10,n()))
```

```
Indiv08_bullers_N<-read.csv("8Indiv_bullers_N.csv") %>% mutate(Individuals = rep(8,n()))
```

```
Indiv06_bullers_N<-read.csv("6Indiv_bullers_N.csv") %>% mutate(Individuals = rep(6,n()))
```

```
Indiv04_bullers_N<-read.csv("4Indiv_bullers_N.csv") %>% mutate(Individuals = rep(4,n()))
```

```
Indiv02_bullers_N<-read.csv("2Indiv_bullers_N.csv") %>% mutate(Individuals = rep(2,n()))
```

```
#combine all observations
```

```
bullers_Ncombined<-
```

```
bind_rows(Indiv02_bullers_N,Indiv04_bullers_N,Indiv06_bullers_N,Indiv08_bullers_N,Indiv10_bullers_N,Indiv12_bullers_N,Indiv14_bullers_N,Indiv16_bullers_N,Indiv18_bullers_N,Indiv20_bullers_N,Indiv22_bullers_N,Indiv24_bullers_N,Indiv26_bullers_N,Indiv28_bullers_N)
```

```
#creating bullers_NSubsets with each of the relevant statistics
```

```
bullers_NcombinedTotSNP<-bullers_Ncombined %>% filter(stat == "Total SNP min. 3") %>% select(TotSNPs = value, Individuals)
```

```
bullers_NcombinedHetSNP<-bullers_Ncombined %>% filter(stat == "Total Heterozygous SNP min. 3") %>% select(HetSNPs = value, Individuals)
```

```
bullers_NcombinedHetENT<-bullers_Ncombined %>% filter(stat == "Total Heterozygous entries min. 3") %>% select(HetENTs = value, Individuals)
```

```
#combining these and then creating an additional column containing the proportion of SNPs
```



```
bullers_NSub<-na.omit(bind_cols(bullers_NcombinedHetSNP, bullers_NcombinedTotSNP,  
bullers_NcombinedHetENT)) %>%
```

```
select(Individuals, HetSNPs, TotSNPs, HetENTs) %>%
```

```
mutate(PropSNPs = HetSNPs / TotSNPs) %>%
```

```
group_by(Individuals) %>%
```

```
filter(HetSNPs>=1000, row_number()>=(n()-999))
```

#Creating an object containing mean value per no. Individuals and a count of observations for each number of individuals (because some observations are lost as NA values during data generation)

```
bullers_NAVG<-bullers_NSub %>%
```

```
group_by(Individuals) %>%
```

```
summarise(n = n(),
```

```
  HetSNPs = mean(HetSNPs),
```

```
  TotSNPs = mean(TotSNPs),
```

```
  PropSNPs = mean(PropSNPs),
```

```
  HetENTs = mean(HetENTs))
```